Uspekhi Mat. Nauk 45:1 (1990), 105–162 Russian Math. Surveys 45:1 (1990), 121–189

Can an individual sequence of zeros and ones be random?

V.A. Uspenskii, A.L. Semenov, and A.Kh. Shen'

CONTENTS

Introduction

If somebody claims that a finite sequence

(1) 000000000000

or

(2) 010101010101

is obtained by tossing a symmetrical coin (0 and 1 correspond to the different sides of the coin), one never believes him. On the other hand, the sequence

(3) 011001011010

does not seem suspicious.

So we definitely have some intuitive notion of a "random sequence" One of the main goals of a mathematical theory is to confirm or to contradict—and therefore to refine—our intuition. (A typical example of the first type is Jordan's theorem: every subset of a plane that is homeomorphic to a circle divides the plane into two parts; famous examples of the second type are mathematical paradoxes.) Can mathematics achieve this goal and give a mathematically rigorous definition of randomness? Naively speaking, the sequences (1) and (2) are not random because their probabilities are too small: 2^{-12}. But the sequence (3) has the same probability! (We shall return to this discussion in the Addendum.)

The problem of randomness for finite sequences (it is better to say "degree of randomness", because a strict distinction between random and non-random finite sequences is hardly possible) was considered in the papers of Kolmogorov and his pupils (see [18], [19], [20], [21], [67], [28], [29], [30], [31], [32], [33], [1],

[2], [3], [4], [57], [58], [59], [60], [61], [62], [63], [65], [66], [51]. Let us mention that this problem is closely connected with the question of theoretical foundations of the Monte-Carlo method.

In our survey we discuss a simpler situation: the sequences are infinite. If we consider infinite sequences as results of the Gedanken experiment—infinite tossing of a symmetrical coin—then all infinite sequences have the same zero probability. Nevertheless, our intuition can distinguish between random and non-random sequences. Now in the case of infinite sequences we can hope to make a strict distinction between random and non-random sequences. Infinite sequences are more abstract objects than finite ones, but the theory of infinite sequences is simpler and can be regarded as a preliminary stage for the theory of randomness of finite objects. From this point of view the notion of an infinite sequence can be regarded as a mathematical model for the notion of "a very long finite sequence". This approach (infinite object as an approximation to a finite object) is typical for mathematics (compare the transitions from the rational numbers to the reals, from the cardinalities of real-world sets to the natural numbers, or from molecular theory to the heat conduction equation).

We restrict ourselves to sequences of zeros and ones (sequences of elements of any finite set can be treated in the same way). We denote by Σ the set of all finite and infinite sequences of zeros and ones. We denote by Ξ the set of all finite sequences of zeros and ones (we call them, as usual, binary words). We denote by Ω the set of all infinite sequences of zeros and ones. So $\Sigma = \Xi \cup \Omega$. In our paper the word "sequence" often means "infinite sequence" (a typical example is the title of the paper). We hope that the reader can distinguish between "sequences" as elements of Ω and "sequences" as elements of Σ from the context.

Our goal is to classify all elements of Ω as "random" or "non-random". Traditional probability theory fails to do this. It never says anything about an individual sequence but only about classes of sequences. When a probability theorist says "Let ω be a random sequence..." and later "ω has a property P" he means only that the property P holds for "almost all" sequences (for all sequences except those which belong to a set having measure zero).

Nevertheless, the problem of a mathematical definition of randomness remains very attractive. The first attempts to solve this problem were made by von Mises [41]. His approach to the definition of a random sequence (Mises uses the word "Kollektiv" instead of "random sequence") is discussed in Ch. VI. We shall discuss there a later development of von Mises' ideas (the so-called "frequency approach" to randomness).

A completely different approach to the notion of random sequence was proposed by A.N. Kolmogorov (see [18], [19],) and developed by Levin ([27]) and Schnorr ([48], [49]). It is called the "complexity approach" and is discussed in Ch. III.

Finally, one of the pupils of Kolmogorov, the Swedish mathematician Per Martin-Löf, developed a quantitative (measure-theoretic) approach to the definition of randomness. In [67] we read: "In 1965 Martin-Löf— using ideas proposed by Kolmogorov—gave a definition of randomness free from these difficulties [connected with the frequency approach of von Mises—Authors' note]. Roughly speaking, Kolmogorov's idea was that "non-random" sequences are sequences having a lot of regularities. Here regularity is a property of a sequence which can be verified and holds for a small part of all sequences (from the measure-theoretic viewpoint)." This quantitative (measure-theoretic) approach to randomness is discussed in Ch. II.

We want to stress that all the three approaches to the definition of randomness lead to a formal mathematical definition by using the theory of algorithms (the notion of computability). We think that one of the main achievements of the theory of algorithms is the definition of randomness (more precisely, definitions of randomness, some of which can—who knows?—pretend to be a "true" definition of randomness). So we can give an affirmative (but probably not final) answer to the question posed in the title of our paper.

As we have said, we devote one chapter to each approach to randomness—frequency approach (von Mises), complexity approach (Kolmogorov) and quantitative approach (Martin-Löf). We try to make these chapters independent of each other as far as possible. We begin with a survey of all three approaches in Ch. I (for the sake of those readers who need only a general survey). So this chapter can be regarded as an introduction to the algorithmic theory of randomness.

Ch. V is devoted to the definition of randomness intermediate between the complexity and the quantitative approaches. This definiton makes use of so-called "probabilistic machines". In Ch. IV we compare the results of the complexity and the quantitative approaches.

We assume that the reader is familiar with the elements of probability theory. So we feel free to say, for example, "measure defined on Borel subsets of Ω" without an explanation. It is a pity that the theory of algorithms, which is another tool important for us, is not included in the standard mathematical curriculum in the Soviet Union. The authors consider this an anachronism. (Let us mention that the theory of algorithms does not coincide with programming; we must also say that many courses given in the technical colleges of the USSR under the titles "discrete mathematics", "theory of algorithms", or "cybernetics", can only discredit each of these subjects.) Nevertheless we must adapt to reality and keep the requirements to a minimum, namely, the reader's familiarity with the theory of algorithms. But "to get used to" is one of the meanings of the word "to understand", so there is no need to explain what "algorithm" means. We hope you realize that some functions are computable (this means that there is an algorithm for computing the value of the function for a given argument). So some elements of Ω

(elements of Ω can be regarded as functions mapping the set of natural numbers into $\{0, 1\}$) are computable. Any computable sequence is non-random according to any further definition of a random sequence of outcomes of a fair coin tossing.

The authors consider the possibility of giving a definition of randomness by means of the theory of algorithms as a strong argument for including the theory of algorithms in the standard curriculum for departments of mathematics in universities.

The problems of the foundations of probability theory have been widely discussed during a long period of time among mathematicians and philosophers. In the Addendum we give a short account of the main lines of this discussion from the viewpoint of the algorithmic theory of randomness.

CHAPTER I

THE MAIN NOTIONS AND FACTS

§1.1. The notion of randomness depends on a given probability distribution

Any attempt to give a rigorous mathematical definition of the notion of randomness must be preceded by a discussion of our intuitive ideas of randomness: otherwise our definition can be a rigorous mathematical definition—but of a different notion.

Here is an evident but essential remark. Let us consider a sequence where there are about twice as many zeros as ones. Can it be random? No, if our coin is symmetric (0 and 1 have equal probabilities); yes, if the coin is not symmetric (one side has probability 1/3, the other side 2/3). We see that the very notion of randomness depends essentially on the probability distribution considered. Up to now we have been in the Bernoulli situation: the sequence of zeros and ones is a sequence of outcomes of independent trials, while in each trial the probabilities of 0 and 1 are p and q $(p + q = 1)$. The more general situation is the Markov one: in this case the probability of 0 and 1 in the n-th trial depends on the result of the preceding trials. This yields another notion of randomness. We may consider even more complicated situations, and each of them will lead to the corresponding notion of randomness.

So the notion of randomness makes sense only with respect to a given probability distribution on Ω. An important class of distributions consists of the Bernoulli distributions—they are determined by probabilities p and q $(p + q = 1)$. The most widely used is the uniform Bernoulli distribution, where $p = q = 1/2$. In this chapter we consider for simplicity only the uniform Bernoulli distribution, because all the essential features can be seen in this case. Before the forthcoming discussion we mention that both the complexity and the quantitative approach can be extended to the more general case of an arbitrary computable probability distribution on Ω (for the exact definition see Ch. II). For the frequency approach it is essential that the distribution is the Bernoulli one (but computability is not necessary). See also the article of Dawid [11] where an attempt is made to give a definition of randomness for non-Bernoulli distributions by means of the frequency approach.

§1.2. Three faces of randomness: stochasticness, chaoticness, typicalness

Random sequences—if they exist at all—have some characteristic properties. We can hope that discussion of these properties can lead us to a formal definition of randomness. For now there is no such definition; we are not able to prove these properties—we can only postulate them as expressing our intuition of randomness (supported by empirical experience and theoretical speculations).

The simplest property of randomness is the frequency stability. In the case of the uniform Bernoulli distribution (the only case discussed in this chapter) this means that the ratio (number of ones in the initial segment of the sequence)/(the segment length) tends to 1/2 as the length tends to infinity. Of course, many non-random sequences also have this property, for example,

$$0101010101010101\ldots.$$

But in this case the subsequence of terms with even numbers does not satisfy the stability property. Therefore, this is not a random sequence: in the case of a true random sequence the stability property holds not only for the sequence itself but also for many of its subsequences (for example, the subsequence formed by the terms with even numbers or the subsequence formed by the terms following 1's). Of course, it is not possible that all subsequences of a given sequence have the stability property. Indeed, if we choose only the terms equal to 0 (or 1), the resulting subsequence contains only zeros (or ones) and hence does not possess the stability property (with respect to the uniform Bernoulli distribution). We can try to impose some restrictions on the class of "allowed" subsequences and call a sequence stochastic if all its "legally chosen" subsequences possess the stability property. Remember our interpretation of a sequence as a sequence of outcomes of a coin tossing: let us imagine that we make a bet, trying to guess the results of coin tossing. Roughly speaking, being stochastic means that there is no winning strategy in the game with the sequence (the terms not included in the chosen subsequence correspond to the coin tossing without a bet). It seems evident that for a random sequence (with respect to the uniform Bernoulli distribution) such a strategy is impossible, so every random sequence is expected to be stochastic.

The basic idea of the frequency approach of von Mises is to identify randomness with stochasticness. Of course, we must give a precise definition of the notion of "allowed selection rule". There are different non-equivalent versions of this definition. They are discussed in Ch. VI, where some disadvantages of the frequency approach are also pointed out.

Another important property of random sequences is chaoticness. This term means that a random sequence is disordered, it has no structure (that is why it is difficult to point to an individual random sequence); the sequence is complex, the simplest way to describe it is to write all its members, and so on. The complexity approach of Kolmogorov consists in identifying chaoticness (complexity) with randomness. Of course, we have to give precise definitions—see §1.4 and Ch. III.

The quantitative approach to randomness is based on the third property of random sequences. This property consists in their typicalness: a random sequence is a typical representative of the class of all sequences. What do we mean by saying that "Mr. Smith is a typical representative of the middle class"? Apparently we mean that: (i) Mr. Smith belongs to the middle class; (ii) he has no specific features (or habits) distinguishing him among the general population of the middle class. In other words, if some feature is specific for a small part of the middle class, Mr. Smith does not have this feature. In a similar way we can say that an infinite sequence ω of zeros and ones is typical if the following property holds: each subset E of the set Ω containing a small part of the whole set Ω does not contain ω. Strictly speaking, this definition has no sense: for each sequence ω the set $\{\omega\}$ contains ω and is an extremely small part of the set Ω. (Each member of the middle class can be considered as non-typical, because people with the same name and date of birth form a very small part of the middle class.) To prevent this difficulty we must restrict ourselves to a specially chosen class of small subsets of Ω. A reasonable class of small subsets was proposed by Martin-Löf. His definition is discussed in this chapter (§1.3 and 1.4).

Of course we can formulate many other properties of random sequences (example: no random sequences are computable) or the class of all random sequences (example: a random sequence remains random after a computable permutation of its terms; more precisely, if $a_n = b_{f(n)}$ for some random b and some computable one-to-one correspondence f between $\mathbb{N}$ and $\mathbb{N}$, then a is random). But we are interested in the "fundamental", "basic" properties of randomness, which imply all others. Maybe typicalness and chaoticness are such properties (they are equivalent, as the Levin – Schnorr theorem says, see §1.4 and Ch. IV). Other natural properties of a random sequence (for example, stochasticness) are consequences of them. From the modern point of view one can say that the essence of randomness is in typicalness and chaoticness. Other properties of a random sequence (among them are different versions of stochasticness) are consequences of these fundamental ones.

§1.3. Typical, chaotic and stochastic sequences: ways to a mathematical definition

We have discussed three properties of randomness (typicalness, chaoticness, stochasticness); sequences possessing these properties are called typical, chaotic and stochastic. Our goal is to find a mathematical definition for these—still vague—notions. When this goal is achieved (if at all), the corresponding precise notions can be regarded as versions of randomness (or some aspects of randomness). In this section we discuss some ways to achieve this goal.

1.3.1. Typicalness

We have already discussed the meaning of the phrase "for all random...". When we say, for example, that "corresponding to the law of large numbers the frequency of ones in a random sequence of zeros and ones tends to 1/2", we mean that the set of all sequences for which this frequency tends to 1/2 has full measure (that is, its complement—the set of sequences such that the frequency of ones has no limit or has limit not equal to 1/2—has measure 0 with respect to the uniform Bernoulli distribution on Ω). It is desirable to define randomness in such a way that the phrase "for all random..." can be understood literally.

Let us formulate our goal more precisely. Let us consider a space X with a probability distribution μ on it. We call sets having measure 0 "null sets". We want—if possible—to define the notion of a random element of X in such a way that for each property P of elements of X the following properties (1) and (2) are equivalent:

(1) almost all elements of X possess the property P (that is, the set of all $x \in X$ that do not possess the property P is a null set);

(2) all random (with respect to μ) elements of X possess the property P.

If our goal is achieved, the phrase "for all random..." can be understood literally (in the example mentioned above $X = \Omega$, μ is an uniform Bernoulli distribution, P is the property "frequency of ones tends to 1/2"). Let us discuss what is needed to reach our goal. Let us assume that we have a definition of randomness such that (1) and (2) are equivalent for all P. Taking the property "to be random" as P we conclude that the set of all non-random sequences is a null set. On the other hand, for each null set U the property "does not belong to U" holds for almost all sequences and must therefore hold for all random sequences. So U must be a subset of the set of all non-random sequences, so the set of all non-random sequences must be the greatest null set up to inclusion.

Therefore our goal cannot be achieved in any non-trivial case: there is no maximal null set (up to inclusion) because all one-element subsets of X are null sets. (This is a more precise version of the argument used in §1.2 to explain why it is necessary to restrict the class of "small subsets of Ω" used in the definition of typicalness.)

Fortunately Martin-Löf (see [37], [38]) discovered that in many cases it is possible to define a subclass of the class of null sets. This class is called the "class of effectively null sets" and has the following properties:

(1) null sets arising in probability theory are usually effectively null sets;

(2) the union of all effectively null sets is an effectively null set. (This set is, therefore, a maximal effectively null set.)

The notion of an effective null set rescues our program: we can define a random element of X as an element that does not belong to any effectively null set (or, in other words, an element that does not belong to the maximal effectively null set). We can regard effectively null sets as "randomness tests": each null set U is a test rejecting all elements of U and a random object is an object passing all tests.

This definition makes the following assertions (1) and (2) equivalent for each property P of elements of X:

(1) the set of elements that do not possess P is an effectively null set;

(2) all random elements of X possess the property P. (This is because the union of all effectively null sets is an effectively null set, as we have mentioned.)

An exact definition of an effectively null set will be given later in this chapter (see §1.4). Chapter II is devoted to a detailed exposition of the quantitative approach to randomness. We have seen that this approach is closely related to the usual practice of probability theory. It is a remarkable fact that it is equivalent to another approach (the complexity one, identifying randomness with chaoticness).

1.3.2. Chaoticness

Kolmogorov used the following way to clarify the notion of a "sequence with a complex structure". At the first stage one defines the complexity measure, which attributes to each binary word a number called the complexity of this word. Then an infinite sequence of zeros and ones is called random if the complexities of its initial segments grow as fast as possible. (The notion of chaoticness in this sense corresponds to randomness with respect to the uniform Bernoulli distribution on Ω. Nevertheless, the complexity approach can be applied to other distributions, see Ch. III.)

What is the difference between simple and complex objects? Why can two sequences of the same length have different complexities? Why does the sequence formed by 1000 zeros look much simpler than a "random" sequence of 1000 zeros and ones? Kolmogorov proposed the following answer: an object is simple if it has a short description. For example, the words "thousand zeros" can be considered as a description of a sequence formed by 1000 zeros. This description is much shorter than the sequence itself, and this is because this sequence is simple. On the contrary, a "random" sequence of 1000 zeros and ones apparently has no simple description, so it is complex. We shall consider binary words (finite sequences of zeros and ones) as descriptions (instead of natural language texts). The size of a description is its length. The complexity of an object (we shall consider binary words as objects) is defined as the length of its shortest description. It remains to define precisely what a "description" is.

It is clear that many different "modes of description" can be invented. So the phrase "x is a description of y" has no sense if we do not fix the mode of description and must be completed by the words "...with respect to a given mode of description". We postpone the exact definition of the word "description" to §1.4. Now we just point out the fact that this definition is given by means of the theory of algorithms.

So we return to the notion of complexity defined as the length of the shortest description. Fortunately Kolmogorov discovered (and this discovery made the theory of complexity of finite objects possible) that among all modes of description one can find so-called "optimal modes of description". There are different optimal modes of description but the corresponding complexities are close to each other (their difference is bounded). Complexity defined with respect to one of the optimal modes of description is called "entropy". After this one can define a random sequence as a sequence such that the entropies of its initial segments increase as fast as possible.

Thus we have given a short outline of Kolmogorov's approach to randomness based on the idea of randomness as the absence of regularities (in the modified terminology). For further discussion see also §1.4 and Ch. III.

1.3.3. Stochasticness

As we have said in §1.2, stochasticness of a sequence means that the property of frequency stability holds for this sequence and for its subsequences obtained by a "legal choice". The outline of different versions of the notion of a "legal choice" (the first of them was proposed by von Mises) is given in the last chapter. Here we restrict ourselves to an example which shows some consequences of frequency stability in legally chosen subsequences of a given sequence. (We assume that some selection rules are legal.)

So we assume that a sequence ω is stochastic and the frequency of ones defined as $p_N =$ (number of ones among the first N terms)/N converges to 1/2 as N tends to infinity. (We recall that we consider only the uniform Bernoulli distribution.) Moreover, the frequency stability holds for all legally chosen subsequences. We assume that among them are all the subsequences defined in the following way. Let X be an arbitrary binary word. Then we can consider a subsequence formed by terms immediately following the occurrences of X. For example, if $X = 01$ and $\omega = 001\underline{0}1\underline{1}101\underline{0}1\underline{1}0\ldots$ then the selected subsequence is formed by the underlined terms. Another example: if $X = 00$ and $\omega = 00000\ldots$ then the third term and all the following terms are included in the selected subsequence.

Considering these selection rules as legal, we require that all subsequences obtained by them have limit of frequencies equal to 1/2.

This requirement implies that all binary words of length k have the same frequency limit (it is equal to 2^{-k}). More precisely, let $\omega = x_0x_1 \ldots$ be a stochastic sequence and S an arbitrary word of length k. Then the ratio (numbers of $i < N$ such that $x_ix_{i+1} \ldots x_{i+k-1} = S)/N$ converges to $1/2^k$ as N tends to infinity.

Let us prove this fact. When $k = 1$ this is an immediate consequence of stochasticness. Arguing by induction we may continue as follows. Let us show that the groups 01 form one fourth of all two-digit groups. Indeed, zeros constitute about a half of all digits. Each zero is followed by zero or by one. These two possibilities are equiprobable (because the sequence formed by the terms following zeros has 1/2 as frequency limit). So the groups 01 form about 1/4 of all two-digit groups. Consider the subsequence formed by the terms following 01. We conclude that the groups 010 and 011 have the same frequency (1/8 of all three-digit sequences), and so on.

So all stochastic sequences are normal in the sense of Borel. According to his definition, a sequence is called normal if "the limits of frequencies ... of digits and combinations of subsequent digits are governed by the same law as in a number with randomly chosen digits, that is, it is equal to 1/10 for one digit, 1/100 for two-digit combinations, 1/1000 for three-digit ones, and so on" (see [6], Russian ed., p. 61; Borel speaks about sequences of decimal digits in the decimal representation of a real number, so he uses 10 instead of 2).

1.3.4. Comments.

Attempts to define the notion of randomness for an individual object were made by von Mises [41], Kolmogorov [18], and Martin-Löf [37] (we list them in chronological order). Von Mises' approach was based on stochasticness: he identified randomness with stochasticness (understood in quite a vague way). He does not use the term "stochastic sequence", calling it "Kollektiv" (see Ch. VI). Kolmogorov's approach was based on chaoticness: he identified randomness with the absence of regularities (we call this "chaoticness"). Kolmogorov did not use the term "chaotic" and spoke about "random sequences". The transfer from finite sequences to infinite sequences (making use of initial segments, this idea was evident to Kolmogorov and his circle) met with some difficulties. These difficulties were overcome simultaneously and independently by Levin (a pupil of Kolmogorov) and Schnorr. Levin and Schnorr invented an adequate version of the definition of entropy (see §1.4.2 and Ch. III). The idea of typicalness is due to the Swedish mathematician Martin-Löf (also a pupil of Kolmogorov); he identified typicalness with randomness (using the term "random" for sequences called typical in our paper).

The definition of a typical sequence given by Martin-Löf (we formulate this definition later in §1.4.1 and study it in Ch. II) was historically the first definition of randomness that is strictly mathematical (unlike von Mises' definition) and adequate (unlike Church's definition discussed in §6.1). When we say "adequate" we mean that it does not contradict our intuition of randomness (as Church's definition does: it leads to a class of random sequences that is evidently too wide). The adequacy of Martin-Löf's definition is confirmed by the fact that it is equivalent to another definition in terms of chaoticness (see §1.4.2 and Ch. IV).

There is another definition of typicalness given by Schnorr. It is considered in §2.3.1. (Like other authors Schnorr did not use the term "typical sequence". These terms ("typical", "chaotic" and "stochastic") were used in our sense publicly for the first time in the opening speech of the First World Congress of the Bernoulli Society on Mathematical Statistics and Probability Theory given by Kolmogorov and Uspenskii (see [23]). Schnorr used the term "zufällig Folge" ([47]) or "Schnorr random sequence" ([49]). Typical sequences (in the sense of Martin-Löf, which we always have in mind unless otherwise stated explicitly) form a proper subclass of the class of all sequences typical in Schnorr's sense. Schnorr called them "hyperzufällig" [47] or "Martin-Löf random" [49].

The notion of typicalness introduced by Schnorr seems to us less adequate to our intuition than Martin-Löf typicalness. This is so for the following three reasons.

(1) Each version of the definition of a typical sequence corresponds to a version of the definition of an effectively null set. There exists a maximal Martin-Löf null set but there does not exist a maximal Schnorr null set (see the statement (1) in §2.3.1). As we have mentioned in §1.3.1, the existence of a maximal null set can be regarded as a merit of the definition of randomness.

(2) Schnorr's definition is not supported by coincidence with chaoticness (Martin-Löf's definition is supported: see the Levin—Schnorr theorem in §1.4).

(3) There is a Schnorr typical sequence that is not stochastic in the sense of Kolmogorov and Loveland (see the Remark in §6.2.1). At the same time our intuition expects that each random sequence is Kolmogorov— Loveland stochastic.

Regarding the von Mises (frequency) approach to randomness we must say that nobody has been able to give a satisfactory frequency definition of randomness up to now. One of the best known attempts is due to Church [10]; another is due to Kolmogorov [17] and Loveland [36]. These definitions are described in §6.1. The Kolmogorov—Loveland definition gives a narrower class than Church's definition, but it is still broader than the class of typical (= Martin-Löf random) sequences (see §6.2.4).

§1.4. Typical and chaotic sequences: basic definitions (for the case of the uniform Bernoulli distribution)

In this section we give exact definitions of typical and chaotic sequences; the proofs (and motivations) are postponed to Chapters II and III. We restrict oursleves to the case of the uniform Bernoulli distribution.

1.4.1. Typicalness.

Let us recall the definition of a null subset of the set Ω (with respect to the uniform Bernoulli distribution). We denote by Ω_x (x is a binary word) the set of all infinite sequences having x as an initial segment:

$$\Omega_x = \{\omega \in \Omega | x \text{ is an initial segment of } \omega\}.$$

We denote by $l(x)$ the length of a binary word x. We call the set $A \subset \Omega$ a *null set* (set having measure 0) if for each $\varepsilon > 0$ there is a sequence (finite or infinite) of binary words x_0, x_1, x_2, ... such that

$$(1) \qquad A \subset \Omega_{x_0} \cup \Omega_{x_1} \cup \dots,$$

$$(2) \qquad 2^{-l(x_0)} + 2^{-l(x_1)} + \dots < \varepsilon.$$

If we use the word "intervals" for sets of the form Ω_x and call $2^{-l(x)}$ the measure of the corresponding interval, we can redefine the notion of a null set in the following way: a set $A \subset \Omega$ is a null set if for each $\varepsilon > 0$ there is a covering of the set A by intervals such that the sum of their measures is less than ε.

Now we define the notion of an effectively null set. (Such sets form a subclass of the class of all null sets.) We obtain the definition of an effectively null set by adding the following additional requirement to the definition of a null set: the sequence of binary words x_0, x_1, ... must be computable and, moreover, the program computing this sequence can be effectively constructed (that is, constructed by an algorithm) from a given $\varepsilon > 0$.

To make this definition precise we must restrict it to rational $\varepsilon > 0$ (instead of all real $\varepsilon > 0$), because (i) otherwise the phrase "effectively constructed from a given $\varepsilon > 0$" is not clear and (ii) for each positive real number $\varepsilon > 0$ there is a smaller positive rational number. So we arrive at the following definition.

Definition of an effectively null (effectively negligible) *set.* A set $A \subset \Omega$ is called an *effectively null set* if there is a computable function $X: \langle \varepsilon, i \rangle \mapsto X(\varepsilon, i)$ (ε is a positive rational number, i is a natural number), the values of X are binary words, and for each $\varepsilon > 0$

$$(1) \qquad A \subset \Omega_{X(\varepsilon,0)} \cup \Omega_{X(\varepsilon,1)} \cup \dots,$$

$$(2) \qquad 2^{-l(X(\varepsilon,0))} + 2^{-l(X(\varepsilon,1))} + \dots < \varepsilon.$$

We do not require that the function X is total. If $X(\varepsilon, i)$ is undefined, then the corresponding term in (1) and (2) must be omitted.

An equivalent definition can be given in terms of enumerable sets. A set is called *enumerable* if there is a program printing all members of this set (this program does not terminate if the set is infinite, but it can go on working infinitely even if the set is finite). Equivalent definition: a set is enumerable if it consists of all values of some computable function. One more equivalent definition: a set is enumerable if it is the domain of a computable function. Using the notion of an enumerable set we can redefine the notion of an effectively null set in the following way: a set $A \subset \Omega$ is an effectively null set if there is an enumerable set W of pairs $\langle$ positive rational number, binary word $\rangle$ such that for each $\varepsilon > 0$

$$A \subset \bigcup \{\Omega_x \mid \langle\varepsilon, x\rangle \in W\}, \tag{1}$$

$$\Sigma \{2^{-l(x)} \mid \langle\varepsilon, x\rangle \in W\} < \varepsilon. \tag{2}$$

Theorem (Martin-Löf [37]). *There is an effectively null set that contains any effectively null set as a subset.*

The proof of this thoerem is given in §2.2. We have explained in §1.3 that this theorem enables us to define a typical sequence as a sequence that does not belong to any null set (= does not belong to the maximal null set). Let us recall once more that we are discussing the case of the uniform Bernoulli distribution on Ω; the general case will be considered in Ch. II.

1.4.2. Chaoticness.

As we said, our first step is to make precise the notion of the "mode of description". Let us recall that we denote by Σ the set of all finite and infinite sequences of zeros and ones. We shall use computable mappings $f: \Sigma \to \Sigma$ as modes of description. Of course, the notion of computability in this case (mappings of Σ into Σ) requires a special definition because arguments and values of f (finite and infinite sequences of zeros and ones) are not constructive objects. We postpone the formal definition to the end of this section. As a "first approximation" the reader may imagine a machine computing the function f in the following sense: $f(x_0 x_1 ...)$ is the sequence of zeros and ones printed on the output when one hits (sequentially) the keys x_0, x_1, ... on the keyboard.

So let us assume that a computable mapping $f: \Sigma \to \Sigma$ is fixed. We call a (finite) binary word y "a description of a (finite) binary word x with respect to a given f" if x is an initial segment of a (finite or infinite) sequence $f(y)$. In other words, to "describe" x means to find a word y such that when it is used as an input for f it generates at the output the word x (maybe, followed by something else).

Definition. The *complexity* of a word x with respect to a computable mapping f (denoted by $KM_f(x)$) is defined as

$$\min\{l(y) | y \text{ is a description of } x \text{ with respect to } f\}.$$

(Here $l(y)$ is the length of the binary word y. As usual, $\min(\varnothing) = +\infty$.)

Definition. A computable mapping $f : \Sigma \to \Sigma$ is called *optimal* if for any computable mapping g there is a constant C such that

$$KMg\ (x) \leqslant KMf\ (x) + C$$

for all binary words x.

Theorem (Kolmogorov). *Optimal mappings do exist.*

Definition. Entropy is a complexity with respect to an optimal computable mapping. So different optimal mappings lead to different notions of entropy. But the difference between any two entropies KM_1 and KM_2 is bounded: the inequality

$$| KM_1\ (x) - KM_2\ (x) | < C$$

holds for some constant C and for all binary words x. So we can say simply "the entropy of x" without indicating a specific optimal mapping (and bearing in mind that entropy is defined only up to a bounded term).

The entropy of a word x is denoted by $KM(x)$.

When the identity mapping ($x \mapsto x$) is used as a mode of description, the complexity of x is equal to its length. Comparing the optimal mode of description with the identity mapping as a mode of description, we conclude that the entropy of an arbitrary word can exceed its length by no more than a constant: $KM(x) \leqslant l(x) + O(1)$. (As usual, $O(1)$ denotes a bounded quantity.)

Sequences such that this inequality becomes an equality for their initial segments are called chaotic. More precisely, we denote by $(\omega)_n$ the initial segment of the infinite sequence ω having length n.

Definition. A sequence ω is called *chaotic* if there is a constant C such that

$$| KM\ ((\omega)_n) - n | \leqslant C$$

for all n.

Theorem (Levin and Schnorr). *The class of chaotic sequences concides with the class of typical sequences.*

Both the proof and references will be given in Chapters III and IV. We conclude this section with the above mentioned formal definition of the computability of a mapping $f : \Sigma \to \Sigma$.

A total mapping $f:\Sigma \to \Sigma$ is called *computable* if the following properties hold:

(1) if a sequence x is an initial segment of a sequence x', then the sequence $f(x)$ is an initial segment of the sequence $f(x')$ (the sequences can be finite or infinite; each sequence is an initial segment of itself);

(2) the value of the mapping f on an infinite sequence x is the minimal sequence that is a continuation of values of f on all finite initial segments of x (according to (1) such a minimal continuation exists);

(3) the set of all pairs $\langle x, y\rangle$ of binary words (finite sequences) such that y is an initial segment of the sequence $f(x)$ is enumerable (by definition an enumerable set is the set of all values of a computable function). The motivation for this definition (including its connection with the informal definition in terms of machines) will be given in Ch. III.

Chapter II

EFFECTIVELY NULL SETS, CONSTRUCTIVE SUPPORT, AND TYPICAL SEQUENCES

This chapter is devoted to the definition of the notion of a typical sequence introduced by Martin-Löf.

§2.1. Effectively null sets, computable distributions, and the statement of Martin-Löf's theorem

As we have noted in Ch. I, a definition of randomness is possible only when the probability distribution is fixed. We shall consider probability distributions on the set Ω of all infinite sequences of zeros and ones. Let us recall that we denote by Ξ the set of all finite sequences of zeros and ones (binary words), while we denote by Ω_x the set of all infinite sequences having x as an initial segment ($x \in \Xi$). The family of sets Ω_x constitutes a basis of a topology on Ω (the standard topology of the Cantor space). So we can speak about Borel subsets of Ω.

As probability distributions we shall consider Borel measures μ on Ω (σ-additive measures defined on the Borel subsets of Ω) such that $\mu(\Omega) = 1$. It is well known that such a measure can be reconstructed from its values on the sets Ω_x for all $x \in \Xi$. Moreover, if $m:\Xi \to \mathbb{R}$ is an arbitrary function from the set of binary words to the reals such that

$$(M)\begin{cases} m(\Lambda)=1 \ (\Lambda \text{ is an empty sequence, that is,} \\ \qquad\qquad\qquad \text{the sequence of zero length}); \\ m(x0)+m(x1)=m(x) \ \text{ for all } x; \\ m(x)\geqslant 0 \ \text{ for all } x, \end{cases}$$

then there is exactly one Borel measure μ such that $\mu(\Omega_x) = m(x)$ for all $x \in \Xi$. We may therefore define the probability distribution on Ω as a function possessing the property (M).

So let us assume that the probability distribution μ on the set Ω is specified. Then the notion of a null subset of Ω (with respect to this distribution) is defined in the usual way. (A null set is not necessarily a Borel set.) In terms of the function $m : x \mapsto \mu(\Omega_x)$ we may define null sets in the following way: a set $A \subset \Omega$ is a null set (= has measure 0 = is negligible) if for all $\varepsilon > 0$ there is a sequence x_0, x_1, ... of elements of Ξ such that

$$A \subset \Omega_{x_0} \cup \Omega_{x_1} \cup \ldots,$$
$$\Sigma m\ (x_i) < \varepsilon.$$

The following facts are well known from classical measure theory. We sketch their proofs in order to be able to compare them with the corresponding effective analogues given later.

(1) If A_0, A_1, ... are null sets, then the set $A_0 \cup A_1 \cup \ldots$ is a null set.

Indeed, to find a covering of the set $\cup A_i$ having the sum of measures less than ε it is enough to find coverings of the sets A_0, A_1, A_2, ... with sums of measures less than $\varepsilon/2$, $\varepsilon/4$, $\varepsilon/8$, ... and take their union.

(2) For each null set A there is a null set B such that $A \subset B$ and B is a G_δ-set (countable intersection of open sets).

Indeed, for each n we choose a covering of the set A by intervals with the sum of measures less than $1/n$. The union of all intervals of this covering is an open set; we denote it by B_n. Then $A \subset B_n$, $\mu(B_n) < 1/n$, and $B = \cap B_n$ will be the set required in (2).

Let us give a definition of an effectively null (= effectively negligible) subset of Ω with respect to the probability distribution μ It is a generalization of the definition given in §1.4 and can be applied to an arbitrary probability distribution μ on Ω (the definition given in the §1.4 is appropriate for the uniform Bernoulli distribution).

Definition. Let μ be an arbitrary probability distribution on Ω and $A \subset \Omega$. We call the set A an *effectively null set* if there is a computable function $X : \langle \varepsilon, i \rangle \mapsto X(\varepsilon, i)$ ($\varepsilon > 0$ is a rational number, i is a natural number, $X(\varepsilon, i)$ is a binary word) such that for all $\varepsilon > 0$

$$(1) \qquad A \subset \Omega_{X(\varepsilon,0)} \cup \Omega_{X(\varepsilon,1)} \cup \ldots,$$

$$(2) \qquad \sum_i \mu\left(\Omega_{X(\varepsilon,i)}\right) < \varepsilon.$$

We do not demand that the function X is a totally defined one. If $X(\varepsilon, i)$ is undefined, the corresponding terms in (1) and (2) are omitted.

Remark. Without loss of generality we may assume that if $X(\varepsilon, i)$ is defined, then $X(\varepsilon, j)$ is defined for all $j < i$. Indeed, we can rearrange the sequence $X(\varepsilon, 0)$, $X(\varepsilon, 1)$, ... so that the terms appear in the same order as they become defined during the parallel computations of $X(\varepsilon, 0)$, $X(\varepsilon, 1)$ ... . Sometimes it is convenient to use this version of the definition of a null set (for example, it is used in [23] and [55]).

This definition can be given with respect to an arbitrary probability distribution μ. However it leads to a reasonable definition of randomness only when the distribution μ satisfies an additional requirement of computability. The notion of computability of a distribution needs a special definition, because the arguments and the values of a measure (Borel sets and real numbers) are non-constructive objects. We therefore restrict μ to the sets Ω_x (for $x \in \Xi$) and require that the values $\mu(\Omega_x)$ can be computed algorithmically with any given precision. So we arrive at the following definition.

Definition. A probability distribution μ is called *computable* if there is a computable function $M: \langle x, \varepsilon \rangle \mapsto M(x, \varepsilon)$ ($x \in \Xi$, ε is a positive rational number, $M(x, \varepsilon)$ is a rational number) such that $|M(x, \varepsilon) - \mu(\Omega_x)| \leqslant \varepsilon$ for all binary words x and for all rational $\varepsilon > 0$.

Theorem (Martin-Löf). *If the probability distribution* μ *is computable, then there is a maximal (up to inclusion) effectively null set (that is, an effectively null set containing all effectively null sets as its subsets).*

Equivalent statement: *the union of all effectively null sets is an effectively null set.*

As we explained in Ch. I, this theorem is a basis for the definition of a typical sequence.

Definition. A sequence is called *typical* (with respect to a given computable probability distribution μ) if it does not belong to any effectively null set (= does not belong to the maximal effectively null set). The set of all typical sequences is called the *constructive support* of the measure μ.

Using this definition we can say that a set is an effectively null set if and only if all its elements are non-typical. The following difference between the classical and the effective case looks like a paradox: the classical definition of a null set restricts the "quantity" of elements in it; the property of "being an effectively null set" depends only on the typicalness of its elements.

Example. A computable sequence ω is typical if and only if the set $\{\omega\}$ has positive measure. (Let us recall that we consider only computable probability distributions.)

Indeed, if $\mu(\{\omega\}) > 0$, then the sequence ω cannot be an element of a null set (and therefore of an effectively null one). So ω is typical. (Here we do not use the computability of the sequence ω. In fact, each sequence ω such that $\mu(\{\omega\}) > 0$ for some computable μ is computable.)

Now let $\mu(\{\omega\}) = 0$. We have to show that $\{\omega\}$ is an effectively null set (and therefore the sequence ω is not typical). Let $(\omega)_n$ be the initial segment of ω having length n. The sets $\Omega_{(\omega)_n}$ decrease as n increases; their intersection is equal to $\{\omega\}$. So their measures converge to 0. Since ω and μ are computable, this convergence is effective: if a rational $\varepsilon > 0$ is given, we can compute algorithmically an initial segment $x(\varepsilon)$ of the sequence ω such that $\mu(\Omega_{x(\varepsilon)}) < \varepsilon$. Now we can use the interval $\Omega_{x(\varepsilon)}$ as a covering of $\{\omega\}$. (Formally: $X(\varepsilon, 0) = x(\varepsilon)$, $X(\varepsilon, k)$ is undefined for $k > 0$.) So $\{\omega\}$ is an effectively null set. Q.E.D.

This example and Martin-Löf's theorem imply that if all one-element sets have measure 0 with respect to a computable probability distribution, then the set of all computable sequences is an effectively null set.

We prove Martin-Löf's theorem in §2.2. We shall discuss now the relation between Martin-Löf's definition of randomness and the usual practice of probability theory. We have already said that the sentence "for a random sequence the property P holds" is an abbreviation for "the set of all sequences for which P does not hold is a null set". This abbreviation can be understood literally if the null set of all sequences for which P does not hold is not only a null set but also an effectively null set. So we come to the following question: can we hope that null sets arising in probability theory are effectively null sets?

Let us consider the law of large numbers as an example. It claims that for almost all sequences $\omega = \omega_0\omega_1 \ldots$ (with respect to the uniform Bernoulli distribution) the equality

$$\lim\, (\omega_0 + \ldots + \omega_{n-1})/n = 1/2$$

holds. In other words, the set of all sequences having limit of frequences not equal to 1/2 or having no limit of frequencies at all is a null set. We have to check that this null set is an effectively null set.

According to Martin-Löf's theorem it is sufficient to prove that for each rational $\varepsilon > 0$ the set S of all sequences $\omega = \omega_0\omega_1 \ldots$ such that

$$|(\omega_0 + \ldots + \omega_{n-1})/n - 1/2| > \varepsilon$$

for infinitely many n is an effectively null set.

We denote by D_n the set of sequences $\omega \in \Omega$ such that

$$|(\omega_0 + \ldots + \omega_{n-1})/n - 1/2| > \varepsilon.$$

We can represent S as $\bigcap_k \bigcup_{n \geqslant k} D_n$ or $\bigcap_k E_k$, where $E_k = \bigcup_{n \geqslant k} D_n$. Now we see that $E_0 \supset E_1 \supset \ldots$ and the law of large numbers (which states that $\mu(S) = 0$) is equivalent to the convergence of $\mu(E_i)$ to 0. (The usual proof of this law uses the upper bound for $\mu(E_i)$ to prove that $\mu(S) = 0$.) These E_i will be the open sets with small measures whose existence is required by the definition of an effectively null set. We must check only that the convergence of $\mu(E_i)$ to 0 is effective (in the ε-N-definition of convergence we can find N effectively from a given ε). This is easy to do by analysing the usual proof of the law of large numbers: in this proof $\mu(D_n)$ is estimated by the use of Stirling's formula for $n!$, and it is shown that $\mu(D_n)$ decreases exponentially as n tends to infinity. (Some details are discussed in §6.2.)

Similar arguments can be applied to other theorems of probability theory.

§2.2. Proof of Martin-Löf's theorem

Roughly speaking, the idea of the proof of Martin-Löf's theorem is the following. We have to prove that the union of all effectively null sets is effectively null. For this purpose we want to use an effective version of the theorem stating that the union of a countable family of null sets is a null set. But we came across a difficulty because there are too many effectively null sets (for example, any subset of an effectively null set is an effectively null set). This difficulty can be avoided by using a special class of effectively null sets (we call sets belonging to this class GN-sets). This class will be chosen in such a way that the following conditions (1) and (2) are satisfied:

(1) every effectively null set is a subset of some GN-set (and all GN-sets are effectively null sets);

(2) the class of all GN-sets is countable and "enumerable" in some sense to be specified later.

Then because of (2) the union of all GN-sets will be an effectively null set; because of (1) this union will be a maximal effectively null set (containing any null set as a subset).

Now we implement this plan. Let us recall that G_δ-sets are countable intersections of open subsets of Ω and that open sets are (countable) unions of intervals (= sets having the form Ω_x for binary words x). So G_δ-sets can be represented in the form

$$\bigcap_i \bigcup_j \Omega_{x(i,j)},$$

where $x(i, j)$ is a binary word (for all natural numbers i, j). By requiring the computability of the function x, we can obtain the effective version of the definition of G_δ-sets. But now we are interested only in "effectively null effectively G_δ-sets" and not even in all such sets. We shall use a class of sets that we call GN-sets. We define a GN-set as a set corresponding to a computable function $x: \langle i, j\rangle \mapsto x(i, j)$ (see above) such that

$$(GN) \qquad \sum_{j<n} \mu\,(\Omega_{x(i,j)}) < \varepsilon$$

for all natural numbers i and n. (As usual we do not require that the function x is total: if $x(i, j)$ is undefined, then the corresponding term is regarded as the empty set.) The requirement (GN) reflects the idea that the measures of sets $\bigcup_j \Omega_{x(i,j)}$ must rapidly decrease as i increases; we formulate our requirement using a finite sum $\sum_{j<n}$ (instead of $\sum_j$) for an important but somewhat technical reason. (Roughly speaking, the reason lies in the necessity to check (*GN*) algorithmically.)

So every set of the form $\bigcap_i \bigcup_j \Omega_{x(i,j)}$, where x is a computable (partial) function satisfying (*GN*), is called a *GN-set* and x is called its *characteristic function*. The definition of a *GN*-set can be reformulated in terms of enumerable sets. Namely, a *GN*-set is a set having the form $\bigcap_i \bigcup_x \{\Omega_x \mid \langle i, x\rangle \in W\}$, where W is an enumerable set of pairs $\langle$natural number, binary word$\rangle$ such that $\mu(\Omega_{x_1}) + \ldots + \mu(\Omega_{x_n}) < 2^{-i}$ for all i and for all $x_1, \ldots, x_n$ such that $\langle i, x_1\rangle \in W, \ldots, \langle i, x_n\rangle \in W$.

***Lemma* 1.** *Any GN-set is an effectively null set. Any effectively null set is a subset of a GN-set.*

Proof. This statement is an immediate consequence of the definitions given above. We must mention only that (i) the values $\varepsilon = 2^{-i}$ are sufficient in the definition of an effectively null set; (ii) although after taking the limit the strict inequality sign $<$ can be replaced by $\leqslant$, this does not matter (we can take a smaller ε).

The next lemma shows that all *GN*-sets can be enumerated.

***Lemma* 2.** *There is a computable (partial) function H of three natural arguments with binary words as values such that*

(1) *for each k the function $\langle i, n\rangle \mapsto H(k, i, n)$ satisfies the requirement (GN) and therefore characterizes the GN-set;*

(2) *each GN-set can be obtained in this way for an appropriate k.*

Proof. We shall use an "effective transformation" which, when applied to an arbitrary computable partial function $x: \mathbb{N} \times \mathbb{N} \to \Xi$, sends it to another computable partial function $y: \mathbb{N} \times \mathbb{N} \to \Xi$. The transformation has the following properties:

(a) each function y obtained by this transformation satisfies the requirement (*GN*);

(b) if a function x satisfies (*GN*), then the function y obtained from it by the transformation coincides with x.

We have to bear in mind that many programs may define the same function (they will be called programs of this function). When speaking about effective transformation of computable functions we mean that an algorithm exists that can be applied to each program of each computable function x and gives a program for some computable function y for which (a) and (b) holds. But different programs for the same function x can be transformed to programs of different functions; this is possible when x does not satisfy (*GN*). (So, strictly speaking, we do not have a transformation of functions but only of programs.)

Having such a transformation we can easily construct the function H required by the statement of the lemma. This can be done by applying this transformation to all programs. More precisely, to compute $H(k, i, n)$ we must take the program number k, apply the transformation to it, and apply the result of the transformation to the pair $\langle i, n \rangle$.

So we have to describe our transformation. An important point is the following remark. If for two real numbers α and β we have algorithms computing approximations to α and β with any given precision and $\alpha < \beta$, then this fact will be discovered sooner or later: the difference between α and β will necessarily become much greater than the approximation errors.

Assume that we have an algorithm computing a function x. We shall describe an algorithm computing the transformed function y. This function is obtained from x by restricting it to some subset of the domain of x. We shall compute simultaneously the values $x(i, 0)$, $x(i, 1)$, ... During this process some computations will terminate. When a new value $x(i, n)$ appears we interrupt our parallel computation and try to check that the new value $x(i, n)$ does not violate the required inequality, that is, the sum of the measures $\mu\ (\Omega_{x(i,\, k)})$ for all k such that $x(i, k)$ is already defined is less than 2^{-i}. We verify this by computing the values of the measure μ with increasing precision. This verification procedure may never terminate if instead of the required inequality we have equality (that is, the sum of measures is equal to 2^{-i}). But if the inequality holds, then it will be discovered and we shall return from the interrupt to the parallel computation of $x(i, 0)$, $x(i, 1)$, ... So if the procedure never terminates, the corresponding function y will have a finite domain. The same is true if the procedure shows that the sum of the measures is greater than 2^{-i} (in this case we abort our computation and the function y has a finite domain). After the procedure is finished with an affirmative answer, we declare that the function y is defined (and is equal to x) on the pairs $\langle i, k \rangle$ for which $x(i, k)$ is already computed.

So in all cases the function y does not violate the requirement (GN); if the given function x itself satisfies this requirement, then y coincides with x. Lemma 2 is proved.

Now we can prove that the union of all GN-sets is an effectively null set. Let $H: \langle k, i, n \rangle \mapsto H(k, i, n)$ be the function from Lemma 2 and let X_0, X_1, ... be the GN-sets obtained from H when $k = 0, 1, \ldots$. To construct a covering of $\bigcup X_n$ by intervals such that the sum of its measures is less than 2^{-i}, we take a covering of X_0 such that the sum of the measures is less than $2^{-i}/2$, a covering of X_1 such that the sum of the measures is less than $2^{-i}/4$, and so on, and then consider the union of all these coverings.

Formally speaking we consider a function z such that $z(i, n) = H(l(n), i + l(n) + 1, r(n))$, where $n \mapsto \langle l(n), r(n) \rangle$ is a one-to-one correspondence between $\mathbb{N}$ and $\mathbb{N} \times \mathbb{N}$. This function z determines a GN-set containing all GN-setsas subsets. Using Lemma 1 we obtain the statement of Martin-Löf's theorem.

§2.3. Different versions of the definition of the notion of typicalness

2.3.1. Schnorr's definition of typicalness.

Schnorr (see [47], [49]) suggested modifying Martin-Löf's definition by imposing some additional requirements of effectively null sets. A set $A \subset \Omega$ is called a *Schnorr effectively null set* if there is a computable function $X : \langle \varepsilon, i \rangle \mapsto X(\varepsilon, i)$ defined for all rational $\varepsilon > 0$ and all natural numbers i (its values are binary words) and also a computable function $N : \langle \varepsilon, \delta \rangle \mapsto N(\varepsilon, \delta)$ defined for all rational $\varepsilon, \delta > 0$ (its values are natural numbers) such that

$$(1) \qquad A \subset \bigcup_i \Omega_{X(\varepsilon, i)} \text{ for all } \varepsilon > 0;$$

$$(2) \qquad \sum_i \mu(\Omega_{X(\varepsilon, i)}) < \varepsilon \text{ for all } \varepsilon > 0;$$

$$(3) \qquad \sum_{i > N(\varepsilon, \delta)} \mu(\Omega_{X(\varepsilon, i)}) < \delta \text{ for all } \varepsilon, \delta > 0.$$

The additional constraint (3) (which is absent in Martin-Löf's definition) implies that the series $\sum_i \mu(\Omega_{X(\varepsilon, i)})$ not only converges to a number less than ε but converges effectively: for each $\delta > 0$ we can effectively find a finite sum such that the difference between it and the sum of the whole series is less than δ. We must mention also that the functions X and N must be total (in Martin-Löf's definition the function X is allowed to be partial). According to this definition the sum $S(\varepsilon) = \sum_i \mu(\Omega_{X(\varepsilon, i)})$ is a computable real number (for all rational $\varepsilon > 0$) and moreover this computability is "uniform" (that is, the program for computing rational approximations to $S(\varepsilon)$ can be obtained algorithmically from a given ε). It is possible to show that by adding this uniform computability requirement to the original definition of an effectively null set (given in §2.1) but still allowing the function X to be partial we obtain a definition equivalent to Schnorr's.

Remark. Here we have modified slightly (and equivalently) Schnorr's original definition of an effectively null set. (Schnorr uses term "total rekursive Nullmenge" for effectively null sets in his sense; see Definition 8.1 in [47].) The changes are made to make the definition simpler and closer to the definition in §2.1. (Schnorr called Martin-Löf effectively null sets "rekursive Nullmenge"; see Definition 4.1 in [47].) In fact Schnorr considers the measure of the union

$$s(\varepsilon) = \mu(\bigcup_i \Omega_{X(\varepsilon, i)})$$

instead of our sum of measures

$$S(\varepsilon) = \sum_i \mu(\Omega_{X(\varepsilon, i)})$$

and requires (instead of (2) and (3)) that

(2′) $s(\varepsilon) < \varepsilon$;

(3′) $s(\varepsilon)$ is an uniformly computable real number.

Schnorr's version of the definition of an effectively null set implies that

(1) for the uniform Bernoulli distribution there is no maximal (up to inclusion) effectively null set (in Schnorr's sense), and the same is true for all the usual probability distributions;

(2) if we define a "Schnorr typical sequence" as a sequence that does not belong to any Schnorr effectively null set, then we obtain a wider class of typical sequences than the class of Martin-Löf typical sequences: there is a Schnorr typical (with respect to the uniform Bernoulli distribution) sequence that is not Martin-Löf typical.

Let us give a sketch of the proofs of the statements (1) and (2). (These facts will not be used later, so we omit the details.) First of all we show that for every Schnorr effectively null set A there is a computable sequence that does not belong to A. (This implies the absence of a maximal Schnorr effectively null set, because the set $\{\omega\}$ is a Schnorr effectively null set for each computable sequence ω.)

When constructing a computable sequence ω that does not belong to a Schnorr effectively null set A we use the existence of a covering of A by intervals with a computable sum of measures that is less than ε for only one value $\varepsilon < 1$. We fix a rational number ε $(0 < \varepsilon < 1)$. Let $x(0), x(1), \ldots$ be a computable sequence of binary words such that $\sum_i \mu(\Omega_{x(i)})$ converges computably and its sum is less than or equal to ε. We shall construct a computable sequence that does not belong to the set $U = \bigcup_i \Omega_{x(i)}$. Let us fix a rational number ε' such that $\varepsilon < \varepsilon' < 1$. We call a binary word x *regular* if the fraction of elements of U among all continuations of x is less than ε':

$$(*) \qquad \mu(U \cap \Omega_x) < \varepsilon' \cdot \mu(\Omega_x).$$

Because of the computable convergence of the series both sides of the inequality $(*)$ can be computed with any given precision. Computing them more and more precisely for all x we can find all regular x. So the set of all regular x is enumerable.

By our assumption the empty word is regular. For each regular x at least one of the words $x0$ and $x1$ is regular. So we can find a computable sequence of regular words such that every succeeding word is obtained from the preceding one by adding 0 or 1 to it. These words are initial segments of a computable sequence of zeros and ones that does not belong to U. Q.E.D.

Let us explain briefly how to construct a Schnorr typical sequence (with respect to the uniform Bernoulli distribution) that is not typical (in the sense used in our paper, that is, in Martin-Löf's sense). We use the coincidence of the class of typical sequences and the class of chaotic sequences (this coincidence will be proved in §4.1). So it is enough to construct a Schnorr typical sequence all initial segments of which have small entropy.

Let us imagine that we want to construct a computable sequence that does not belong to a Schnorr effectively null set A. Consider a covering with a computable sum of measures less than ε' (we denote the union of these intervals by U) and construct a sequence of regular words of increasing length (initial segments of a sequence $\omega \in A$). Assume that at some stage of this construction we decide to add a new requirement: the sequence ω must not belong to some other Schnorr effectively null set B. At this stage we have a regular word x, that is, a word x such that $\mu(U \cap \Omega_x) < \varepsilon' \cdot \mu(\Omega_x)$. If a set V is a union of intervals covering B with sufficiently small sum of measures, then the word x will be regular with respect to $U \cup V$ (that is, $\mu((U \cup V) \cap$ $\cap \Omega_x) < \varepsilon' \cdot \mu(\Omega_x))$. Choosing such a V we can continue the construction of regular words— now regular with respect to $U \cup V$. This construction gives a computable sequence ω that does not belong to $A \cup B$. If at some stage of this construction we decide that ω must not belong to a third Schnorr effectively null set C, we can find a covering W of the set C having sufficiently small measure such that the current regular word remains regular when we add W, and so on.

One may suppose that such a construction can give a computable sequence that does not belong to any Schnorr effectively null set (if we consider subsequently all pairs of computable functions X and N satisfying conditions (2) and (3) of the definition of a Schnorr effectively null set). Nevertheless this is not so, since this construction requires additional information on whether the pair X, N satisfies the requirements (2) and (3) or not, so we are not able to construct a computable sequence. Nevertheless, if we take a new pair of functions X, N into consideration at a sufficiently late stage of our construction, then the amount of this additional information will be small compared with the length of the already constructed part of the sequence. In this case the entropy of initial segments of the sequence will be small compared with their lengths, that is, this sequence is not chaotic.

This argument can be considered as a sketch of a proof of the following result: there is a Schnorr typical (with respect to the uniform Bernoulli distribution) sequence such that the entropy of its initial segment of length n does not exceed $C.\log_2 n$ for all n and for some C (independent of n).

2.3.2. Solovay's criterion for typicalness.

Another variant of the definition of randomness is proposed in [8], [9] and is called there "R.M. Solovay randomness". (In Chaitin's papers random real numbers are considered, but all definitions are valid mutatis mutandis for sequences of zeros and ones.)

An infinite sequence ω is called "Solovay random" with respect to a computable probability distribution μ if there is no computable partial function $X : \mathbb{N} \to \Xi$ such that $\sigma = \sum_i \mu(\Omega_{X(i)}) < \infty$ and $\omega \in \Omega_{X(i)}$ for infinitely many i.

Clearly a non-typical sequence ω is not Solovay random: if we take coverings of the set $\{\omega\}$ having measures less than 1, 1/2, 1/4, ... then their union is a covering having measure less than 2 such that ω belongs to infinitely many intervals of this covering. The reverse implication is also true: a sequence that is not Solovay random is not typical. This can be proved by a simple argument attributed in [9], p. 36 to Solovay. Let ω belong to $\Omega_{X(i)}$ for infinitely many i. Let us denote by U_n the set of all sequences belonging to $\Omega_{X(i)}$ for at least n different values of i. Then $\omega \in U_n$ for all n. It is easy to show that each U_n can be represented as a union of a computable family of non-intersecting intervals and that $\mu(U_n) \leqslant \sigma/n$. So we can effectively construct a covering of the set $\{\omega\}$ having arbitrary small measure. (Possible non-computability of the real number σ does not contradict the effectiveness of this construction, because we can use any rational number larger than σ instead of σ.)

So Solovay randomness is equivalent to typicalness (=Martin-Löf randomness).

2.3.3. The axiomatic approach to the definition of typicalness.

This approach differs radically from the algorithmic one; its idea is due to J. Myhill. Let us consider typicalness as a new undefined notion and formulate axioms of typicalness expressing our intuition. For example, we can add to second-order arithmetic (or set theory) a new predicate symbol $R(\omega)$; we read $R(\omega)$ as "the sequence ω is typical". The axiom scheme reflecting our intuition of typicalness may be formulated as follows:

$$\forall \omega\ (R\ (\omega) \Rightarrow \varphi\ (\omega)) \Leftrightarrow (\mu\ (\{\omega \mid \neg\ \varphi\ (\omega)\}) = 0)$$

(μ is the probability distribution defined on Ω which we consider; it is easy to show that the right-hand side of this equivalence can be expressed in second-order arithmetic). If we allow an arbitrary formula φ in this axiom scheme, this theory will become inconsistent. If we require that $\varphi(\omega)$ has no free variables (except ω) and does not contain R, then this theory becomes consistent but uninteresting. We do not know whether this theory will be consistent if we require that $\varphi(\omega)$ has no free variables other than R in φ. It is easy to show that such a theory is inconsistent with the axiom of constructibility $V = L$.

The axiomatic approach to the definition of typicalness and stochasticness (see Ch. VI) is discussed in a recent paper by Lambalgen [26].

CHAPTER III

COMPLEXITY, ENTROPY, AND CHAOTIC SEQUENCES

In this chapter we give an exposition of the complexity approach to randomness based on the natural idea that randomness is an absence of regularities. It became possible to make this idea precise when Kolmogorov [18] introduced the notion of the entropy of a finite object. (The notion that we call entropy is called complexity in [18]; the term "entropy" in our sense was used for the first time in [19].) After this the natural idea arose that randomness of an infinite sequence can be defined in terms of the entropies of its initial segments.

However, this plan encountered some difficulties (see [37], and also, for example, [16] and [67]). These difficulties were overcome in 1973 by Levin [27] and Schnorr [48]. To make the definition of randomness of an infinite sequence in terms of entropies of its initial segments possible it was necessary to use so-called monotone entropy instead of the simple Kolmogorov entropy defined in [18]. Using this notion we may call a sequence chaotic if the entropies of its initial segments grow as fast as possible (see details later) and obtain a class of chaotic sequences equal to the class of typical sequences (in the sense of Martin-Löf, see Ch. II). This coincidence gives the motivation for considering chaoticness as a formalization of the intuitive notion of a "random" sequence.

§3.1. Computable mappings

Let us give a definition of a computable mapping of the set Σ of all finite and infinte sequences into itself. The natural way to do this is to use general constructions of the theory of f_0-spaces in the sense of Ershov [12] (see [53] for the application of the notion of f_0-space to definitions of entropy). We do not consider the general case, but give the definition of computability only for mappings of Σ into Σ. We give two versions: first using a relatively abstract language and then its interpretation in a more concrete way. (The reader is free to choose between these two versions.)

Let us introduce a partial ordering on $\Sigma : x \leqslant y$ if the sequence x is a prefix (initial segment) of the sequence y (x and y may be infinite). The sequences x and y are called *comparable* if $x \leqslant y$ or $y \leqslant x$. For each finite sequence x the set of all finite and infinite continuations of x (sequences having x as a prefix) is denoted by Σ_x. Let us consider the family Σ_x as a base of a topology on the space Σ (open sets are unions of the sets having the form Σ_x). Let us mention that the space Σ is not a Hausdorff space (T_2-space). It is not even a T_1-space, but only a T_0-space. All computable mappings (according to our definition) are continuous (with respect to this topology) total functions mapping Σ into Σ. It is easy to show that the continuity of a total function $f : \Sigma \to \Sigma$ is equivalent to the conjunction of the following two conditions:

(a) if $x, y \in \Sigma$, $x \leqslant y$, then $f(x) \leqslant f(y)$;

(b) the value of f on an infinite sequence x is the least upper bound of the values of f on its finite prefixes: $f(x) = \sup\{f(x_0) | x_0 \text{ is finite}, x_0 \leqslant x\}$. (Note that the condition (a) guarantees that the least upper bound in (b) does exist.) So any continuous mapping $f : \Sigma \to \Sigma$ is completely determined by its values on finite sequences. For each continuous mapping f we can consider the set F of all pairs $\langle p, q\rangle$ of finite sequences such that $q \leqslant f(p)$. A continuous mapping f can be reconstructed when F is known:

$$f(x) = \sup \{q \mid \exists p \, (p \leqslant x \text{ and } \langle p, q\rangle \in F)\}.$$

We call the function f and the set F *conjugate*. This conjugacy relation is a one-to-one correspondence between continuous total mappings $f : \Sigma \to \Sigma$ and subsets $F \subset \Xi \times \Xi$ such that for all p, q, p', q', q_1, q_2

(1) $\langle p, \Lambda\rangle \in F$

(2) $\langle p, q\rangle \in F,\ p' \geqslant p,\ q' \leqslant q \Rightarrow \langle p', q'\rangle \in F$;

(3) $\langle p, q_1\rangle \in F,\ \langle p, q_2\rangle \in F \Rightarrow q_1$ and q_2 are comparable.

We call a continuous mapping $f : \Sigma \to \Sigma$ *computable* if the set F conjugate to f is enumerable (=is the set of all values of a computable function = is the set of pairs being printed during the execution of a program).

So the definition of a computable function is the following. Let F be the enumerable set of pairs of binary words (=finite sequences of zeros and ones) possessing the properties (1)–(3). Let $f : \Sigma \to \Sigma$ be a function defined by the formula

$$f(x) = \sup \{q \mid \exists p \, (p \leqslant x \text{ and } \langle p, q\rangle \in F)\}.$$

All functions obtained in this way (they are continuous) are called *computable mappings* of Σ into Σ. (We stress that according to this definition computable mappings are totally defined on Σ; instead of undefined values of f they have values equal to Λ.)

Another (more concrete) definition of a computable mapping of Σ into Σ can be given as follows. Imagine a computer having input and output. The input is a sequence of zeros and ones (one may imagine a user hitting the keys "0" and "1" on the keyboard). The output is a sequence of zeros and ones appearing on the printer. A computer of this type gets a (finite or infinite) sequence as its input and prints a (finite or infinite) sequence as its output. (We consider a non-terminating execution of a program; it is allowed to continue computations while waiting for the next input symbol.)

In general the output sequence of zeros and ones depends not only on the input sequence but also on the moments of appearance of input symbols. However, we shall consider only programs such that the output sequence depends only on the input sequence (but not on the moments of their appearance). Such a program defines a mapping of Σ into itself (input sequence $\mapsto$ output sequence). These mappings are called computable mappings of Σ into itself.

For example, the program may simply ignore the input and send successive binary digits of the number π to the output. This means that the constant mapping equal to the binary representation of π on all inputs is computable.

(In general, the constant mapping is computable if and only if its value is a computable sequence of zeros and ones.) Another example of a computable mapping is the identity mapping: the program copies the input to the output.

The programs used in this definition are versions of the "oracle" algorithms (see [46], §9.2).

§3.2. Kolmogorov's theorem. Monotone entropy

The notion of monotone entropy was introduced by Levin (see [27]) and (independently and simultaneously but in a slightly different way) by Schnorr [48]. (Later Schnorr [49] rejected his first version and used Levin's notion from [27].) This notion is a variant of the notion of entropy of a finite object proposed by Kolmogorov. We will not discuss the original definition of Kolmogorov (see [18], [21], [55]) but give only the definition of monotone entropy introduced by Levin.

Let $f : \Sigma \to \Sigma$ be a computable totally defined mapping. Let y be an arbitrary element of Σ. If $y \leqslant f(x)$ for some $x \in \Xi$, then x will be called a *description* of y. The *complexity* of y with respect to f is defined by the formula

$$\inf \{l(x) \mid x \text{ is a description of } y\} = \inf \{l(x) \mid y \leqslant f(x)\}.$$

Here $l(x)$ denotes the length of x; as usual, $\inf(\varnothing) = +\infty$. We denote the complexity of y with respect to f by $KM_f(y)$. We say that a mapping f is no worse than a mapping g if $KM_f(y) \leqslant KM_g(y) + O(1)$, that is, there is a constant C such that $KM_f(y) \leqslant KM_g(y) + C$ for all $y \in \Sigma$ (notation: $KM_f \underset{+}{\underset{\frown}{<}} KM_g$).

Theorem (Kolmogorov). *Optimal* (*that is, no worse than any other*) *mappings exist.*

Proof. The proof is based on a very simple idea. Let us consider all computable mappings of Σ into itself. They form a countable set, so we assign natural numbers to them. The optimal mapping f can be constructed as follows: if x is a description of y with respect to the n-th mapping, then the pair $\langle n, x \rangle$ is a description of y with respect to f. In other words, the value of f on $\langle n, x \rangle$ is equal to the value of the n-th mapping on x.

Why does this construction lead to an optimal mapping? Let $g : \Sigma \to \Sigma$ be an arbitrary computable mapping. Let n be its number. If x is a description of y with respect to g, then the pair $\langle n, x \rangle$ is the description of the same y with respect to f. So the amount of additional information required to transfer from g to f (that is, the number n) depends only upon g (and does not depend upon y).

This argument must be made more precise in the following way:

(1) we must enumerate all computable mappings of Σ into Σ in such a way that a mapping f will be computable;

(2) the mapping f must be defined on binary words (not pairs), so we need some coding of pairs by words;

(3) this coding must satisfy the following requirement: the length of the code of a pair $\langle n, x\rangle$ must exceed the length of x by no more than a constant (independent of x).

Such specifications are possible. To enumerate all computable mappings of Σ into Σ we use the "universal algorithm" (an interpreter which applies any program to any input data). The pair $\langle n, x\rangle$ (n is a natural number, x is a binary word) can be coded by the word 0^n1x (here 0^n is the word containing n zeros). It is possible to reconstruct both n and x from 0^n1x and the length of the code exceeds the length of x by $n+1$.

Now we give a formal construction of the required optimal mapping (first by using the definition of computability in terms of sets of pairs and then by using programs with inputs and outputs).

Using sets of pairs. We need an enumerable set W of triples $\langle n, p, q\rangle$ (n is a natural number, p and q are binary words) such that the following statements hold:

(1) for all n the set $W_n = \{\langle p, q\rangle \mid \langle n, p, q\rangle \in W\}$ is conjugate to a computable mapping of Σ into Σ (that is, it satisfies the requirements (1)–(3) of the definition of a computable mapping in §3.1);

(2) among W_n one can find all sets of pairs satisfying the requirements (1)–(3).

(The existence of such a set is often expressed by the phrase "the family of all computable mappings from Σ to Σ is enumerable".)

Let us assume that such a W exists. Then the mapping conjugate to the set of pairs

$$U = \{\langle 0^n1p, q\rangle \mid \langle n, p, q\rangle \in W\} \cup \{\langle p, \Lambda\rangle \mid p \in \Xi\}$$

is optimal. First of all we must show that U is conjugate to some computable mapping (that is, U satisfies the requirements (1)–(3)). This is so because all W_n satisfy these requirements and because 0^n1q can be the prefix of 0^m1p only if $m = n$ and q is the prefix of p.

We denote the mapping conjugate to U by f. It is optimal: if g is an arbitrary computable mapping conjugate to a set G having a number n ($W_n = G$), then $KM_f(y) \leqslant KM_g(y)+n+1$ for all $y \in \Sigma$. Indeed, if x is a description of y with respect to g (that is, $y \leqslant g(x)$), then $\langle x, y\rangle \in G$, $\langle n, x, y\rangle \in W$, $\langle 0^n1x, y\rangle \in U$, so 0^n1x is a description of y with respect to f. Its length exceeds the length of x by $n+1$. (We assume that y is finite; if y is infinite, then the same argument can be applied to any finite initial segment of y.)

So it is sufficient to construct a set W with the required properties. It is well known that there is a so-called universal enumerable set V of triples, that is, an enumerable set V such that among the sets

$$V_n = \{\langle p, q\rangle \mid \langle n, p, q\rangle \in V\}$$

(called sections of V) all enumerable sets of pairs can be found. Evidently, some of V are "bad": they do not satisfy the requirements (1) – (3) of the definition of a computable mapping. Such V_n are not conjugate to any computable mappings. We shall transform V into the set W in such a way that all sections become "good" and all "good" sections remain unchanged.

This transformation is made as follows: when enumerating elements of V we "delete inconsistencies" and "fill gaps". "Deleting inconsistencies" means that if an element $\langle n, p, q\rangle$ appears in the enumeration of V and it contradicts an element $\langle m, r, s\rangle$ already included in W (in the sense that $m = n$, r is comparable with p, and q is not comparable with s—we recall that two words are comparable if one is a prefix of the other), then $\langle n, p, q\rangle$ has to be "deleted" (it is not included in the enumeration of W). "Filling gaps", on the contrary, means adding to W some new elements which do not belong to V. Namely, for each triple $\langle n, p, q\rangle$ included in W we add to W all triples $\langle n, p', q'\rangle$ such that $p' \geqslant p$, $q' \leqslant q$.

It is easy to see that the transformation described makes all sections "good" and does not change good sections.

Using programs with inputs and outputs. We enumerate all programs of this type. The optimal program will act as follows. The program waits for the appearance of the digit "1" on the input and counts the zeros preceding it. After this (if the input contains 1 at all) the program imitates the n-th program (where n is the number of zeros preceding the first 1) on the input formed by the digits following the first 1. So if x is a description of y with respect to the n-th program, then 0^n1x is a description of y with respect to the above mentioned optimal program. Note that the length of the description with respect to the optimal program is not greater than the length of the description with respect to the n-th program plus $n+1$. Q.E.D.

However, this description of an optimal program has a serious defect. We recall that our programs must satisfy the correctness requirement, which states that the output sequence depends only on the input sequence (but not on the moments of time when the terms of the input sequence appear). To make our optimal program correct in this sense we must imitate only correct programs. If the programming language used does not guarantee correctness, we must distinguish between correct and incorrect programs; this is impossible to do effectively for any natural programming language. So our construction looks impossible. However, it can be corrected by using a class K of programs such that:

(1) it is possible to decide algorithmically whether a given program belongs to K;

(2) all programs in K are correct;

(3) for each correct program there is an equivalent program in K (two programs are called equivalent if the corresponding mappings of Σ into Σ coincide).

If such a class K exists, then we can imitate only programs in K. In fact, it exists, but its construction is no easier than considering sets of pairs. We omit this construction: the reader may either believe us or be satisfied with our first proof (in terms of sets of pairs).

Let f and g be optimal mappings. The complexities KM_f and KM_g differ only by a bounded additive term: $KM_f(x) = KM_g(x) + O(1)$. It is convenient to fix an optimal mapping f, call $KM_f(x)$ the monotone entropy of the sequence x, and denote it by $KM(x)$, omitting f. However, we must remember that the fixed optimal mapping f can be chosen arbitrarily. So actually the function KM is defined up to a bounded additive term.

Although our definition can be applied to infinite sequences (and the entropy of an infinite sequence is finite if and only if it is computable) we shall use in the sequel only entropies of finite sequences (binary words). We shall omit the word "monotone" because we do not use other variants of the notion of entropy here.

Remark. In [23] and [55] a slightly different definition of monotone entropy is given. It is equivalent to our definition in the sense that the difference between these two entropies is bounded by a constant. In [23] and [55] a *mode of description* is defined as an enumerable relation $R \subset \Xi \times \Xi$ such that if $\langle x_1, y_1 \rangle \in R$, $\langle x_2, y_2 \rangle \in R$, $x_1 \leqslant x_2$, then y_1 and y_2 are comparable. Then the complexity of a word y with respect to a given mode of description is defined as $\min\{l(x) | \langle x, y \rangle \in R\}$, and entropy is defined as the complexity with respect to an optimal mode of description (a mode of description is called optimal if the corresponding entropy is minimal up to an additive constant). This definition is equivalent to ours because

(1) if f is a computable mapping of Σ into Σ, then the set of pairs conjugate to f is a mode of description in the sense of [23];

(2) if R is a mode of description in the sense of [23], then the mapping $f : \Sigma \to \Sigma$ defined by the formula

$$f(u) = \sup \{y \in \Xi \mid (\exists x \in \Xi)\ (\langle x, y\rangle \in R \text{ and } x \leqslant u)\}$$

is computable.

§3.3. Chaotic sequences

In this section we generalize the definition of a chaotic sequence (given in Ch. I for the case of the uniform Bernoulli distribution). This generalized definition can be applied to any computable distribution on Ω. (We call a probability distribution on Ω *computable* if there is an algorithm that computes approximations to $\mu(\Omega_x)$ with any given precision. Let us recall that Ω_x is the set of all infinite continuations of a binary word x.)

It is natural to call a sequence chaotic if the sequence of entropies of its initial segments grows as fast as possible. Of course the words "as fast as possible" require a precise formulation.

Lemma. *Let P be a computable probability distribution on Ω. Then a constant C exists such that for any binary word x the following inequality holds:* $KM(x) \leqslant -\log_2 P(\Omega_x) + C$.

This lemma motivates the following definition.

Definition. Let P be a computable probability distribution on Ω. A sequence $\omega \in \Omega$ is called *chaotic* with respect to P if the set

$$\{(-\log_2 P(\Omega_x)) - KM(x) \mid x \text{ is an initial segment of } \omega\}$$

is bounded.

In other words, a chaotic sequence is a sequence such that for its initial segments x the inequality $KM(x) \leqslant -\log_2 P(\Omega_x) + O(1)$ (see the lemma) becomes an equality.

Proof of the lemma. For each binary word x we define a segment V_x on the real line in such a way that the length of V_x is equal to $P(\Omega_x)$, $V_\Lambda = [0, 1]$ (Λ is an empty word), V_{x0} and V_{x1} are two parts of V_x separated by some point (V_{x0} is the left part, V_{x1} is the right part). The correspondence $x \mapsto V_x$ is uniquely determined by these requirements. Note that it can be constructed for each probability distribution on Ω. For the uniform Bernoulli distribution this construction gives a family $\{I_x\}$ of segments such that I_x contains numbers with binary representation in Ω_x (we omit evident reservations concerning endpoints of segments). These segments will be used in our proof together with the segments V_x corresponding to the probability distribution P.

Let us define a computable mapping f by the requirement that y is a description of a non-empty word x with respect to f if the segment I_y is contained in the interior of the segment V_x. (We consider an interior with respect to $[0, 1]$, so 0 and 1 are internal points of $[0, 1]$.) In other words, $f: \Sigma \to \Sigma$ is conjugate to the set of pairs

$$F = \{\langle y, x\rangle | x = \Lambda \text{ or } (I_y \subset (\text{the interior of } V_x))\}.$$

This set of pairs is enumerable (=is the set of all values of a computable function). Indeed, the endpoints of the segments V_x can be computed algorithmically with any given precision (because P is computable). So if I_y is contained in the interior of V_x, then this fact can be established at some stage of the computation. (Here it is essential that we use the interior of V_x: if the endpoints of I_y and V_x coincide, any precision is not sufficient to discover that I_y is contained in V_x.) Computing the endpoints of all V_x with increasing precision and taking all pairs $\langle y, x\rangle$ such that we know already that I_y is contained in the interior of V_x, we obtain the enumeration of F. It is easy to show that F is conjugate to a computable mapping (that is, it satisfies the requirements (1)–(3) of §3.1).

It remains to prove that

$$KM_f(x) \leqslant -\log_2 P(\Omega_x) + O(1)$$

(for the mapping f constructed above). Indeed, if $-\log_2 P(\Omega_x)$ is less than or equal to a natural number n, then the length of V_x is not less than 2^{-n}. So the interior of V_x (like the interior of any segment having this length) contains a segment I_y (for some y) having length $(1/4)\cdot 2^{-n}$. This y is a description of x with respect to f and $l(y) \leqslant n+2$ and, therefore, $KM_f(x) \leqslant n+2$. So $KM_f(x) \leqslant -\log_2 P(\Omega_x)+3$ (here 3 appears instead of 2 because the logarithm is not necessarily an integer). The lemma is proved.

We have mentioned that the classes of chaotic and typical sequences coincide for any computable distribution P. The proof of this statement is given in the next chapter.

CHAPTER IV

WHAT IS A RANDOM SEQUENCE?

In this chapter we prove that for each computable probability distribution on the set Ω the classes of typical (Ch. II) and chaotic (Ch. III) sequences coincide. This fact was proved in 1973 independently by Levin [27] and Schnorr ([48], [49]). (As we have mentioned in §3.2, Schnorr used in his first paper [48] another version of entropy but this difference has practically no influence on the proofs. In [49] he defines and uses the notion of entropy equivalent to Levin's discussed in Ch. III.)

§4.1. The proof of the Levin–Schnorr theorem for the uniform Bernoulli distribution

Theorem (Levin and Schnorr). *A sequence is typical with respect to the uniform Bernoulli distribution on Ω if and only if it is chaotic* (*with respect to the same distribution*).

Proof. According to the definitions of typical and the chaotic sequences we must prove the following assertions:

(1) if the difference $n-KM((\omega)_n)$ is unbounded, then the set $\{\omega\}$ is an effectively null set;

(2) if $\{\omega\}$ is an effectively null set, then the difference $n-KM((\omega)_n)$ is unbounded. (We denote by $(\omega)_n$ the initial segment of ω having length n.)

Proof of the assertion (1). We have seen in §1.4 that the difference $n-KM((\omega)_n)$ can be unbounded only from above. Let us assume that $n-KM((\omega)_n)$ is unbounded from above. Then for each c there is an initial segment x of the sequence ω such that $KM(x) < l(x)-c$. (Here $l(x)$ denotes the length of the word x.) We denote by D_c the set of all binary words x such that

$KM(x) < l(x)-c$. As we have seen, for each c there is an initial segment of the sequence ω that belongs to D_c. We can consider binary words as vertices of a binary tree (its root is an empty word). An infinite sequence of zeros and ones can be considered as a path in this tree starting from the root. The path corresponding to ω intersects each set D_c (for any c). To prove that ω is not typical we must prove (in accordance with the definition) that $\{\omega\}$ is an effectively null set. This means that we can effectively find (from a given ε) a set (a union of intervals) having measure less than ε and containing ω.

We can use for this purpose the set P_c of all sequences having an initial segment in D_c. Indeed, this set contains ω. Let us prove that the measure of P_c does not exceed 2^{-c} (with respect to the uniform Bernoulli distribution). Let x_0, x_1, ... be all minimal elements of D_c (elements of D_c such that their prefixes different from themselves do not belong to D_c). Evidently $P_c = \bigcup \Omega_{x_i}$ and, therefore, it is sufficient to prove that $\sum 2^{-l(x_i)} \leqslant 2^{-c}$. We know that $l(x_i) \geqslant KM(x_i)+c$, so it is sufficient to prove that $\sum 2^{-KM(x_i)} \leqslant 1$. The latter inequality is a consequence of the following simple lemma (its proof is postponed to §4.3).

***Lemma* 1.** *Let x_0, x_1, ... be incomparable binary words (for all $i \neq j$ the word x_i is not a prefix of x_j). Then $\sum 2^{-KM(x_i)} \leqslant 1$.*

So for every $\varepsilon > 0$ we can find a covering of the set $\{\omega\}$ by intervals with the sum of measures less than ε. The definition of an effectively null set requires that this covering is given in the form $\Omega_{X(\varepsilon,0)}$, $\Omega_{X(\varepsilon,1)}$, . . . where X is a computable function such that $\sum P(\Omega_{X(\varepsilon,i)}) < \varepsilon$. Comparing what is desired with what has so far been obtained we conclude that it is natural to use the elements of D_c (for sufficiently large c such that $2^{-c} < \varepsilon$) as $X(\varepsilon,0)$, $X(\varepsilon, 1)$, ... The computability of the function X is a consequence of the following lemma (for the proof see §4.3).

***Lemma* 2.** *There is an algorithm enumerating all $\langle c, x\rangle$ such that $KM(x) < l(x)-c$.*

Now only one problem remains: the definition of an effectively null set requires that the sum of the measures of the intervals covering it is less than ε, and we have proved that the measure of the union of the intervals is less than ε. This problem cannot be solved by taking only minimal elements of D_c, because we need computability. We must use the following lemma.

***Lemma* 3.** *For each computable sequence x_0, x_1, ... we can effectively construct a computable sequence y_0, y_1, ... of binary words such that any two words y_i, y_j $(i \neq j)$ are incomparable and $\bigcup\Omega_{x_i} = \bigcup\Omega_{y_i}$.* (The word "effectively" means that a program computing x_i from a given i can be effectively transformed to a program computing y_i from a given i.)

This lemma is proved in §4.3. Its application finishes the proof of the assertion "typical $\Rightarrow$ chaotic" of the Levin–Schnorr theorem. We shall now prove another assertion ("chaotic $\Rightarrow$ typical").

Let ω be a non-typical sequence. We must prove that the difference $n-KM((\omega)_n)$ is unbounded. According to the definition of typicalness our non-typical sequence ω is an element of an effectively null set. So we can effectively find a covering of the set $\{\omega\}$ by intervals with arbitrarily small sum of measures. We must use the existence of such a covering to find initial segments of ω having small entropies. The idea of this argument can be explained as follows. If we know that ω belongs to a set M having small measure, then we can construct a computable mapping adapted to the set M, that is, such that the elements of M have short descriptions (but others have long descriptions or have no descriptions at all). Of course, the preceding sentence should not be understood literally (for example, the elements of M are infinite sequences but entropy is defined for finite sequences). The precise formulation is given in the following lemma.

***Lemma* 4.** *Let A be an effectively null set. Then for each c we can effectively produce a computable mapping $f:\Sigma\to\Sigma$ such that the following property holds: each sequence $\omega\in A$ has an initial segment x such that $KM_f(x)<l(x)-c$.* ("Effectively" means that there is an algorithm computing from a given c a program enumerating the set of pairs conjugate to the mapping f.)

It seems that this lemma cannot be used to finish the proof of the Levin–Schnorr theorem, because different values of c correspond to different mappings f. But we can easily cope with this difficulty. Let us consider a computable sequence $c_0, c_1, \ldots$ of natural numbers that grows fast enough. Using Lemma 4 we construct the corresponding mappings $f_0, f_1, \ldots$. We combine them into one computable mapping f as we did in the proof of Kolmogorov's theorem. Namely, we define f by the formula $f(0^n1x) = f_n(x)$. Then $KM_f(y) \leqslant KM_{f_n}(y)+n+1$. So if $\omega\in A$, then for each n there is an initial segment x of the sequence ω such that $KM_{f_n}(x) < l(x)-c_n$ and, therefore, $KM_f(x) \leqslant KM_{f_n}(x)+n+1 < l(x)-c_n+n+1$. By an appropriate choice of c_n (it is enough to use $c_n = 2n$) we can prove that $l(x)-KM_f(x)$ is unbounded from above for initial segments x of the sequence ω. So $l(x)-KM(x)$ is unbounded from above.

It remains to prove Lemma 4. According to the definition of an effectively null set, for each $\varepsilon > 0$ we can effectively find a computable sequence of binary words $x_0, x_1, \ldots$ such that $\sum P(\Omega_{x_i}) < \varepsilon$ and $A \subset \bigcup \Omega_{x_i}$. Let us recall that a sequence $x_0, x_1, \ldots$ can be chosen in such a way that it has no gaps (if x_i is defined, then x_j is defined for $j < i$; we mentioned this fact in §2.1). Each sequence $\omega\in A$ has an initial segment among the x_i. So our goal will be achieved if we construct a computable mapping f such that $KM_f(x_i) < l(x_i)-c$ for all i. To do this we choose (in a way to be described later) incomparable words $y_0, y_1, \ldots$ and then choosing f such that y_i is a description of x_i with respect f. (In terms of sets of pairs: consider a set $\{\langle p, q\rangle | q = \Lambda \text{ or } \exists i\ (y_i \text{ is a prefix of } p \text{ and } q \text{ is a prefix of } x_i)\}$ and a mapping f conjugate to this set.) Then $KM_f(x_i) \leqslant l(y_i)$ for this f, so it is sufficient to choose y_i in such a way that $l(y_i) < l(x_i) - c$.

So our goal can be described as follows: we have a sequence n_i of natural numbers ($n_i = l(x_i) - c$); we need a sequence of binary words $y_0, y_1, \ldots$ such that $l(y_i) = n_i$ and y_i is incomparable with y_j for all $i \neq j$.

The necessary condition for the existence of such y_i is $\sum 2^{-n_i} \leqslant 1$ (the sum of the measures of disjoint sets Ω_{y_i} does not exceed 1). This condition is in fact sufficient, but in order to simplify the proof we shall use the stronger restriction $\sum 2^{-n_i} \leqslant 1/2$.

***Lemma* 5.** *Let n_i be a computable sequence of natural numbers and $\sum 2^{-n_i} \leqslant 1/2$. Then there is a computable sequence of binary words y_i such that $l(y_i) \leqslant n_i$.*

This lemma will be proved in §4.3. By using it we are able to complete the proof of Lemma 4 (and the proof of the Levin–Schnorr theorem) by establishing the inequality

$$\sum 2^{-l(x_i)+c} = 2^c \cdot \sum 2^{-l(x_i)} \leqslant 1/2.$$

Because $\sum 2^{-l(x_i)} = \sum P(\Omega_{x_i})$ is less than ε (by our assumption) it is sufficient to take ε small enough (for example, $\varepsilon < 1/2^{c+1}$).

So the Levin–Schnorr theorem is proved for the uniform Bernoulli distribution. In the next section we discuss the changes necessary for the case of an arbitrary computable probability distribution. The proofs of Lemmas 1, 2, 3, 5 are given in §4.3.

The proof of the Levin–Schnorr theorem leads to the following interesting corollary. We have proved that if ω is not typical, then $n - KM((\omega)_n)$ is unbounded. Just the same argument shows that if ω is not typical, then $n - KM((\omega)_n) \to \infty$ as $n \to \infty$. To obtain this conclusion we must modify the proof of Lemma 4. In this proof the word y_i was a description of x_i with respect to f. Now we define the mapping f in another way and assume that $y_i z$ is a description of $x_i z$ for each word z. For this f the inequality $KM_f(x) \leqslant l(x) - c$ holds not only for $x = x_i$ but for all x having some x_i as an initial segment.

So for typical ω the difference $n - KM((\omega)_n)$ is bounded and for non-typical ω the difference tends to infinity as n tends to infinity. So there is no sequence ω such that this difference is unbounded but does not tend to infinity.

§4.2. The case of an arbitrary probability distribution

The proof of the Levin–Schnorr theorem given in §4.1 can easily be adapted to the case of an arbitrary probability distribution. Let us exhibit briefly the necessary changes.

1. We used the inequality $KM(x) \leqslant l(x)+O(1)$. Now we need the inequality $KM(x) \leqslant -\log_2 P(\Omega_x)+O(1)$ instead (see §3.3).

2. The set D_c must be defined as the set of words x such that $KM(x) < -\log_2 P(\Omega_x)-c$. As before, it is enumerable (here we make use of the computability of the probability distribution P). The set of all continuations of all its elements has measure less than 2^{-c} (with respect to the distribution P).

After these changes are made the first part of the proof of the Levin–Schnorr theorem goes as before. In the second part of the proof we use the following generalization of Lemma 4.

***Lemma* 4a.** *Let A be an effectively null set (with respect to the distribution P). Then for each c a computable mapping $f : \Sigma \to \Sigma$ can be effectively obtained such that the following property holds: each sequence $\omega \in A$ has an initial segment x such that $KM_f(x) < -\log_2 P(\Omega_x)-c$.* ("Effectively" means that there is an algorithm computing from a given c a program enumerating the set of pairs conjugate to the mapping f.)

The proof is similar to the proof of Lemma 4 given in §4.1. We must define the numbers n_i as $\lfloor -\log_2 P(\Omega_{x_i}) \rfloor - c$ (instead of $n_i = l(x_i)-c$; $\lfloor z \rfloor$ means the integer part of z). After this the proof of the Levin–Schnorr theorem is completed as in §4.1.

§4.3. The proofs of the lemmas

***Lemma* 1.** *Let $x_0, x_1, \ldots$ be incomparable binary words (for all $i \neq j$ the word x_i is not a prefix of x_j). Then $\Sigma 2^{-KM(x_i)} \leqslant 1$.*

Proof. Let f be the optimal computable mapping used in the definition of KM. Let y_i be the shortest description of the word x_i. Any two words y_i, y_j $(i \neq j)$ are incomparable. (If a word y is a common continuation of y_i and y_j, then $f(y)$ is a common continuation of x_i and x_j, which does not exist by hypothesis.) So the sets Ω_{y_i} are disjoint and the sum of their measures (with respect to the uniform Bernoulli distribution) does not exceed 1. The measure of Ω_{y_i} is equal to $2^{-l(y_i)}$, that is, it is equal to $2^{-KM(x_i)}$ because y_i is the shortest description of x_i and $l(y_i) = KM(x_i)$. Lemma 1 is proved.

***Lemma* 2.** *There is an algorithm enumerating all $\langle c, x \rangle$ such that $KM(x) < l(x)-c$.*

Proof. Let f be the optimal computable mapping and F the set of pairs of words conjugate to f:

$$F = \{\langle p, x \rangle | p, x \text{ are binary words, } x \text{ is a prefix of } f(p)\}.$$

According to the definition of a computable mapping of Σ into Σ the set F is enumerable, that is, there is an algorithm enumerating all pairs $\langle p, x \rangle \in F$. For each pair $\langle p, x \rangle \in F$ and for each c we check whether the inequality $l(p) < l(x)-c$ holds. If this is the case, then the inequality $KM(x) < l(x)-c$ holds for the pair $\langle c, x \rangle$ and we can include this pair in the enumeration of the set of pairs mentioned in the lemma. Each pair of this set will appear in the course of this process, since for any given x we can find a word p such that $\langle p, x\rangle \in F$ and $l(p) = KM(x)$. Lemma 2 is proved.

***Lemma* 3.** *For each computable sequence $x_0, x_1, \ldots$ we can effectively construct a computable sequence $y_0, y_1, \ldots$ of binary words such that any two words y_i, y_j $(i \neq j)$* are incomparable and $\bigcup \Omega_{x_i} = \bigcup \Omega_{y_i}$. (The word "effectively" means that a program computing x_i from a given i can be effectively transformed to a program computing y_i from a given i.)

Proof. Sets having the form Ω_x (x is a binary word) are called *intervals*. During this proof we call a subset of Ω regular if it is a finite union of intervals. It is easy to see that each regular set can be represented as a union of disjoint intervals and that the set-theoretic difference of two regular sets is a regular set. (For example, we can prove this fact by using the following remark: a set A is regular if and only if there is a natural number n such that for any sequence we can decide whether it is contained in A by using only its prefix of length n.)

Now it is easy to describe the transformation of the sequence $x_0, x_1, \ldots$ into the sequence $y_0, y_1, \ldots$. Assume that the initial segment $x_0, x_1, \ldots, x_k$ is transformed to $y_0, y_1, \ldots, y_l$, where the words $y_0, \ldots, y_l$ are incomparable and

$$(*) \qquad \Omega_{x_0} \cup \ldots \cup \Omega_{x_k} = \Omega_{y_0} \cup \ldots \cup \Omega_{y_l} .$$

After receiving the next element x_{k+1} we have to add to the right-hand side of the equality (*) some disjoint intervals such that their union is equal to

$$(\Omega_{x_0} \cup \ldots \cup \Omega_{x_{k+1}}) \setminus (\Omega_{x_0} \cup \ldots \cup \Omega_{x_k}).$$

This can be done because the difference is a regular set. Obviously all transformations described above preserve the computability of the sequence. Lemma 3 is proved.

***Lemma* 5.** *Let n_i be a computable sequence of natural numbers and $\sum 2^{-n_i} \leqslant 1/2$. Then there is a computable sequence of binary words y_i such that $l(y_i) \leqslant n_i$.*

Proof. As in §3.3 we consider the correspondence $x \mapsto I_x$ between binary words and segments of the real line such that I_x contains real numbers with binary representation in Ω_x (again evident reservations concerning the endpoints are omitted). We call all the segments I_x *regular*. Besides these we use segments $S_0, S_1, \ldots$ such that the length of S_i is equal to $2 \cdot 2^{-n_i}$, the left endpoint of S_0 is 0, and the left endpoint of S_{i+1} coincides with the right endpoint of S_i. By our assumption all these segments S_i are included in the segment [0, 1]. For each S_i we consider the largest regular segment I_{y_i} included in S. Each segment having length l contains a regular segment having length greater than or equal to $l/2$. Therefore the length of I_{y_i} is not less than 2^{-n_i}, which means that $l(y_i) \leqslant n_i$. Moreover, the segments I_{y_i} with different i are disjoint (because the segments S_i are disjoint). Therefore the words y_i are incomparable. Lemma 5 is proved.

CHAPTER V

PROBABILISTIC MACHINES, A PRIORI PROBABILITY, AND RANDOMNESS

§5.1. Probabilistic machines

The notion of a probabilistic machine reflects the idea of a computer having a "random number generator" (RNG). Such a machine has two parts: a RNG and a deterministic part processing the information obtained from the RNG.

We assume that the information obtained from the RNG has the form of an infinite sequence of zeros and ones. Then the RNG can be characterized by the probability distribution on its outcomes, that is, by the measure on the space Ω of all infinite sequences of zeros and ones (the measure of the whole space Ω is equal to 1).

Exploring the possibilities of probabilistic machines, one may impose some restrictions on the RNG. (Otherwise it can happen, for example, that some sequence ω of zeros and ones forms a set of full measure. This sequence may contain arbitrary information and so, roughly speaking, probabilistic machines may do everything.)

The natural requirement is the computability of the probability distribution on the outcomes of the RNG. (Let us recall that computability means that the probability of the event "a binary word x is a prefix of the outcome of the RNG" can be computed with any given precision for a given x, see §2.1.) The simplest RNG is a symmetric coin (zeros and ones are equiprobable and trials are independent). Such a generator corresponds to the uniform Bernoulli distribution on the set Ω. We shall see that in some sense this generator is enough: any RNG with a computable probability distribution can be "simulated" using the symmetric coin.

Let us give precise definitions. We define a *probabilistic machine* as a pair $\langle P, f \rangle$, where P is a computable probability distribution on Ω and f is a computable mapping of Σ into Σ. Let us explain the informal meaning of this pair: P is the probability distribution on the outcomes of the RNG; the transformation f is applied to a sequence obtained from the generator. The value of f is the output of a probabilistic machine. (Here we consider probabilistic machines not as a tool for computing functions but as a tool for generating sequences of zeros and ones, so they have no input except that of the RNG.) So each probabilistic machine determines a random variable with values in the set Σ of all finite and infinite sequences of zeros and ones, and we can consider its distribution.

Let $\langle P, f \rangle$ be the probabilistic machine. We can define the probability distribution Q on Σ associated with this machine as follows. The measure $Q(A)$ of an arbitrary Borel subset $A \subset \Sigma$ will be defined as $P(f^{-1}(A))$. (Strictly speaking, we should write $P(f^{-1}(A) \cap \Omega)$ instead of $P(f^{-1}(A))$, because the set $f^{-1}(A)$ can contain finite sequences. For brevity we neglect this difference and identify the probability distribution P defined on Ω with the probability distribution defined on Σ which coincides with P on Ω and is equal to zero on Ξ.) Let us mention that while the distribution P is equal to zero outside Ω, the distribution Q does not necessarily have this property, because the mapping f can have a finite value for an infinite argument. The distribution Q defined in this way is called the *distribution associated with the probabilistic machine* $\langle P, f \rangle$.

The natural question arises: which distributions on Σ are associated with probabilistic machines? It appears that these distributions have a simple description. Let us recall that we denote by Σ_x the set of all finite and infinite continuations of a binary word x. For any probability distribution Q on the set Σ (defined on the Borel subsets of Σ) the following properties hold:

(1) $Q\ (\Sigma_\Lambda) = 1$ (the measure of the whole set Σ is equal to 1);

(2) $Q\ (\Sigma_{x0}) + Q\ (\Sigma_{x1}) \leqslant Q\ (\Sigma_x)$ for all binary words x.

The inequality (2) is not an equality if the Q-measure of the set $\{x\}$ differs from zero. The standard measure-theoretic argument shows that a distribution on Σ is determined by its values on sets having the form Σ_x. The only requirement on these values is that conditions (1) and (2) hold: if q is an arbitrary function on binary words with non-negative real values such that $q(\Lambda) = 1$ and $q(x0)+q(x1) \leqslant q(x)$ for all binary words x, then there is a unique probability distribution Q on Σ such that $Q(\Sigma_x) = q(x)$ for all x.

So our question can be formulated as follows: which conditions on the function q are necessary and sufficient for the existence of a probabilistic machine such that Q is associated with it? To formulate such conditions we need to introduce the notion of a "real number semicomputable from below". A real number x is called *semicomputable from below* if there is a computable increasing sequence of rational numbers converging to x. Any computable number is semicomputable from below. Indeed, if x_n is its rational approximation having precision $1/n$, y_n is equal to $x_n - 1/n$, and $z_n = \max(y_1, ..., y_n)$, then the sequence is an increasing computable sequence of rational numbers converging to x. Let us mention equivalent definitions of a real number x semicomputable from below: "x is the least upper bound of an enumerable set of rational numbers", "the set of all rational numbers less than x is enumerable". The notion of a real number semicomputable from above can be defined in a similar way. If a real number x is semicomputable from above and from below, then x is computable. Indeed, when enumerating approximations from above and from below converging to x we can wait until the difference between approximations from above and from below becomes arbitrarily small.

A function q defined on all binary words having real values is called *semicomputable from below* if all its values $q(x)$ are semicomputable (from below) real numbers and corresponding programs can be obtained effectively from a given x. More precisely (and in slightly different terms) q is semicomputable from below if there is a computable function $\langle x, n\rangle \mapsto \bar{q}(x,n)$ with rational values defined for all words x and all natural numbers n such that for any x the sequence $\bar{q}(x, 0)$, $\bar{q}(x, 1)$ ... is an increasing sequence converging to $q(x)$.

An equivalent definition in terms of enumerable sets: a function q is enumerable from below if the set of pairs

$$\{\langle x, r\rangle | x \text{ is a binary word, } r \text{ is a rational number, } r < q(x)\}$$

is enumerable.

As the following theorem states, semicomputability from below is a necessary and sufficient condition for the existence of a probabilistic machine.

Theorem. *Let Q be an arbitrary probability distribution on Σ and let $q(x) = Q(\Sigma_x)$ for all binary words x. Then the following properties are equivalent:*

(1) *there is a probabilistic machine such that Q is associated with it;*

(2) *there is a probabilistic machine $\langle P, f\rangle$ such that Q is associated with it and P is a uniform Bernoulli distribution;*

(3) *the function q is semicomputable from below.*

Proof. Evidently, (2) implies (1). So it is sufficient to prove that (1) $\Rightarrow$ (3) and (3) $\Rightarrow$ (2). (We have already mentioned the equivalence (1) $\Leftrightarrow$ (2) when saying that the coin as a RNG is sufficient to simulate any RNG with a computable probability distribution.)

Let us prove that (1) implies (3), that is, for any probabilistic machine $\langle P, f\rangle$ the function $q : x \mapsto P(f^{-1}(\Sigma_x))$ is semicomputable from below. Indeed, the preimage $f^{-1}(\Sigma_x)$ can be represented as the union of sets Σ_y for all $y \in \Xi$ that are descriptions of x with respect to f : $f^{-1}(\Sigma_x) = \bigcup\{\Sigma_y | x \leqslant f(y)\}$. The set of all such y is enumerable and can be represented as $\{y_0, y_1, ...\}$, where y_i form a computable sequence. Then

$$q\,(x) = \lim P\,(\Omega_{y_0} \bigcup \ldots \bigcup \Omega_{y_n}) \quad \text{as} \quad n \to \infty.$$

The real number $P\,(\Omega_{y_0} \bigcup \ldots \bigcup \Omega_{y_n})$ is computable: we choose y_i having no prefixes among y_0, ..., y_n except themselves and compute the sum of their measures (here we use the fact that the sum of computable real numbers is computable). So $q(x)$ can be represented as the limit of an increasing sequence of computable real numbers; it is easy to transform it into an increasing computable sequence of rational numbers. (Take approximations from below with sufficient accuracy and make the sequence increasing by replacing the n-th term by the maximum of the first n terms.) So the real number $q(x)$ is semicomputable from below. All our constructions can be carried out effectively when x is given, so the function q is semicomputable from below. The implication (1) $\Rightarrow$ (3) is proved.

The proof of the implication (3) $\Rightarrow$ (2) is more subtle. We call a subset of the space Σ *effectively open* if it can be represented in the form $\bigcup\{\Sigma_s | s \in S\}$, where S is an enumerable set of binary words. A program enumerating the set S will be called a program of the corresponding effectively open set. (Of course, an effectively open set may have different programs.) If f is a computable mapping of Σ into itself and $T_y = f^{-1}(\Sigma_y)$, then the following statements hold:

(a) for each $y \in \Xi$ the set T_y is effectively open, and its program (one of its programs) can be obtained algorithmically from a given y;

(b) $T_\Lambda = \Sigma$, and for each y the sets T_{y0} and T_{y1} are disjoint subsets of the set T_y.

Conversely, if for each y the set T_y is given such that conditions (a) and (b) are satisfied, then there is exactly one computable mapping of the set Σ into itself such that $f^{-1}(\Sigma_y) = T_y$ for all y.

If $\langle P, f \rangle$ is a probabilistic machine and T_y is a family of effectively open sets corresponding to f, then the distribution Q associated with the probabilistic machine $\langle P, f \rangle$ can be described as follows: $Q(\Sigma_y) = P(T_y)$. So we must prove that for any function q semicomputable from below there is a family of sets satisfying (a) and (b) such that $P(T_y) = q(y)$ for all y.

These sets T_y can be constructed by induction on the length of y. First of all we put $T_\Lambda = \Sigma$. The induction step uses the following lemma.

Lemma. *Let X be an effectively open set and r, s two non-negative real numbers semicomputable from below such that $r+s \leqslant P(X)$. Then we can effectively find two effectively open sets Y, Z that are disjoint subsets of X and such that $P(Y) = r$, $P(Z) = s$.* ("Effectively" means that there is an algorithm using programs for X and programs computing increasing rational approximations to r and s as its inputs, and giving programs for Y and Z as output.)

By applying this lemma to the set T_Λ and the real numbers $q(0)$, $q(1)$ we obtain the sets T_0 and T_1. Then, by applying the lemma to the set T_0 and the real numbers $q(00)$, $q(01)$, we obtain the sets T_{00} and T_{01}, and so on (the sets T_{x0} and T_{x1} are obtained by applying the lemma to the set T_x and the real numbers $q(x0)$, $q(x1)$).

A sketch of the proof of the lemma. We shall construct effectively open sets Y and Z as follows. The set X is represented as the union of the set Σ_x for all x from some enumerable set S. We shall "distribute" these sets Σ_x between Y and Z. We know that the real numbers r and s are computable from below; consider increasing computable sequences of rational approximations converging to r and s. Their terms will be considered sequentially and called "current approximations" to r and s. Receiving an interval from the set X, we distribute it between Y and Z in such a way that:

(1) the measures of the parts of Y and Z already constructed do not exceed the current approximations to r and s;

(2) if possible, these measures are equal to the current approximations (if possible means that the measure of the intervals of X considered up to now is sufficient).

When the interval is fully distributed we proceed to the next interval. It is easy to see that the sets Y and Z constructed as described above satisfy the requirements of the lemma. The lemma is proved.

So we have a characterization of images of computable measures under computable mappings. The functions q defined on all binary words taking non-negative real values such that $q(\Lambda) = 1$, $q(x0)+q(x1) \leqslant q(x)$ are often called *semimeasures* and identified with the corresponding distributions on Σ. Distributions associated with probabilistic machines correspond to semimeasures that are semicomputable from below.

§5.2. A priori probability

As we have seen in §5.1, for each probabilistic machine an associated probability distribution on Σ can be constructed. The following theorem states that among these distributions there exists a maximal one.

Theorem. *There is a probability distribution M on Σ associated with a probabilistic machine such that for each distribution Q associated with the probabilistic machine we can find a constant c such that $Q(A) \leqslant cM(A)$ for any Borel set $A \subseteq \Sigma$.* (Notation: $Q \overset{\cdot}{\lesssim} M$.)

Proof. Theorem 5.1 implies that it is sufficient to consider only probabilistic machines with a symmetric coin as a random number generator (that is, pairs $\langle P, f \rangle$, where P is a uniform Bernoulli distribution). Let us consider a probabilistic machine working as follows. First of all it chooses at random a natural number n. This can be done in an arbitrary way; it is only necessary that the probabilities of all natural numbers differ from zero. (For example, n can be the number of heads preceding the first tail.) Then our probabilistic machine simulates the probabilistic machine $\langle$symmetric coin, n-th computable mapping$\rangle$: all the following digits obtained using the coin are considered as a sequence, and the n-th computable mapping is applied to this sequence. (We discussed the numbering of all computable mappings in §3.2 above.)

We denote by M the probability distribution associated with this machine. Let Q be a probability distribution associated with an arbitrary probabilistic machine. We may assume that this machine is $\langle P, f \rangle$, where P is a uniform Bernoulli distribution on Ω, and f is a computable mapping of Σ into itself. Then the probability p of an event "universal machine simulates the machine $\langle P, f \rangle$" is positive: p is greater than or equal to the probability of the event "choosing randomly a natural number n we obtain a number of the mapping f". So $M(A) \geqslant p \cdot Q(A)$ and $Q(A) \leqslant (1/p) \cdot M(A)$. Q.E.D.

Speaking more formally, we consider a probabilistic machine $\langle P, f \rangle$, where P is the uniform Bernoulli distribution and f is a universal mapping of

Σ into Σ used in the proof of the Kolmogorov theorem in §3.2. This f was constructed in such a way that $f(0^n1x)$ is equal to the value of the n-th computable mapping on x. Therefore $f^{-1}(A)$ contains all sequences $0^n1\alpha$ such that α belongs to the preimage of A with respect to the n-th computable mapping of Σ into Σ. Hence, $M(A) \geqslant 2^{-n-1}\cdot Q(A)$, where n is the number of the mapping corresponding to the distribution Q. The theorem is proved.

Corollary. *There is a semimeasure semicomputable from below that is maximal up to a constant factor among all semimeasures semicomputable from below.*

Let us note that this corollary gives only part of the information contained in the theorem because it concerns only measures of sets having the form Σ_x (in general, there is no way to transfer from these sets to an arbitrary set). This corollary can be proved directly. Namely, all semimeasures semicomputable from below can be enumerated (we enumerate all functions semicomputable from below and convert them into semimeasures in such a way that semimeasures remain unchanged). Then we consider computable series Σp_i, where all $p_i > 0$ and their sum is equal to 1, and take a function $m = \Sigma p_n \cdot$ (the n-th semimeasure semicomputable from below).

The probabilistic machine constructed in the proof of the theorem can be called "universal" in the following sense: if some probabilistic machine gives an element of a set $A \subset \Sigma$ with a positive probability, then this "universal" machine also gives an element of A with a positive probability.

If M_1 and M_2 are two probability distributions on Σ for which the statement of the theorem is valid, then $M_1(A) \leqslant c_1M_2(A)$ and $M_2(A) \leqslant c_2M_1(A)$ for some constants c_1, c_2 and all (Borel) sets A. So M_1 and M_2 differ only by a factor which is bounded and separated from zero.

Let us specify some probability distribution M for which the statement of the theorem is true. We call it "a priori probability".

§5.3. A priori probability and entropy

We have defined two characteristics of a binary word x: its monotone entropy $KM(x)$ and the a priori probability $M(\Sigma_x)$ of the set of all its continuations; the latter will be denoted by $m(x)$. We can say that $KM(x)$ measures the difficulty of the task "describe x or its continuation" and that $m(x)$ shows the probability of an accidental appearance of x. It turns out that these two characteristics are closely connected.

Theorem. *The following inequalities hold*:

$$-\log_2 m(x) \leqslant KM(x) + O(1); \tag{1}$$

$$KM(x) \leqslant -\log_2 m(x) + O(\log_2 l(x)). \tag{2}$$

Proof. The first inequality is quite evident. Indeed, let f be a computable mapping used in the definition of the monotone entropy. Let us consider the probabilistic machine $\langle P, f \rangle$, where P is a uniform Bernoulli distribution on Ω. If $KM(x) = n$, then there is a binary word y of length n such that x is a prefix of $f(y)$. In this case for any infinite continuation $\omega \in \Omega$ of the binary word y the sequence $f(\omega)$ belongs to Σ_x. So the set $f^{-1}(\Sigma_x)$ contains all continuations of y and its measure is not less than 2^{-n}. It remains to use the maximal property of the a priori probability and take logarithms.

The second inequality is more complicated. Roughly speaking, we can describe the difficulty as follows: the set $f^{-1}(\Sigma_x)$ (f is a computable mapping used in the definition of the a priori probability) can be broken into small pieces Σ_t, where all the t are relatively long but due to a large number of different t's the union $\bigcup \Sigma_t$ has a relatively large measure. This argument shows that we cannot expect that $KM(x) \leqslant -\log_2 m(x) + O(1)$ (a counter-example to this inequality was constructed by Gács in [14]) and we are forced to add $O(\log l(x))$ to the right-hand side of this inequality. Let us mention also that the "typical" value (typical for the majority of words having given length) of $KM(x)$ (and $-\log_2 m(x)$) is equal to $l(x)$ in order of magnitude, so this logarithmic term is usually small compared with the others.

The main part of the proof is contained in the following lemma. A sequence $p_0, p_1, \ldots$ of real numbers is called *semicomputable from below* if there is a computable function $p : \langle i, n \rangle \mapsto p(i, n)$ defined on all pairs of natural numbers such that for each i the sequence $p(i, 0), p(i, 1), \ldots$ increases and converges to p_i.

Lemma. *Let $p_0, p_1, \ldots$ be a semicomputable from below sequence of non-negative real numbers such that $\Sigma p_i < \infty$. Let $x_0, x_1, \ldots$ be an arbitrary computable sequence of binary words. Then $KM(x_i) \leqslant -\log_2 p_i + O(1)$.*

(Let us mention that the word x_i has nothing in common with the number p_i except for their number i!)

Before proving the lemma we show that the inequality (2) of the theorem is a consequence of the lemma. For each binary word x we consider the real number $m(x)/l(x)^2$. The series $\sum_x m(x)/l(x)^2$ converges. Indeed, the sum of all $m(x)$ for all x having fixed length n cannot exceed 1, because these x are incomparable and the corresponding Σ_x are disjoint. So $\sum_x m(x)/l(x)^2 \leqslant \sum_n 1/n^2$ (we group together terms corresponding to words of equal length). All binary words can be arranged in a computable sequence $x_0, x_1, \ldots$; let p_i be equal to $m(x_i)/l(x_i)^2$. The sequence p_i is semicomputable from below (because the function m is semicomputable from below). Using the lemma we obtain $KM(x_i) \leqslant -\log_2(m(x_i)/l(x_i)^2) + O(1) \leqslant -\log_2 m(x_i) + 2 \log_2 l(x_i) + O(1)$. Q.E.D.

Proof of the lemma. Assume that the numbers p_i have the form 2^{-n_i}, where $n_0, n_1, \ldots$ is a computable sequence of natural numbers. We may assume without loss of generality that Σp_i does not exceed 1/2 (a constant factor in p_i can be compensated by $O(1)$). Let us recall Lemma 5 from §4.3; by using it we can construct a computable sequence of incomparable binary words y_0, y_1, ... such that $l(y_i) = n$. Now we consider a computable mapping $f : \Sigma \to \Sigma$ such that $f(y_i) = x_i$ and find that $KM_f(x_i) \leqslant l(y_i) = n_i = -\log_2 p_i$, so $KM(x_i) \leqslant$ $\leqslant -\log_2 p_i + O(1)$.

Let the numbers p_i be uniformly computable (that is, the program computing approximations to p_i can be obtained effectively from a given i). Then we can replace the p_i with their approximations having the form $1/2^k$ and differing from p_i by no more than a factor 2, and finally reduce the problem to the preceding case.

Let us now consider the general case: the sequence p_i is semicomputable from below (it is just the case necessary for the proof of inequality (2) of the theorem). Here we can use the following trick. When computing the approximations to p_i from below, we look for inequalities of the form $1/2^k < p_i$ (for all k and i) that are guaranteed by the already known approximations. All numbers $1/2^k$ discovered during this process form a computable sequence. The sum of all these numbers is less than or equal to $2\Sigma p_i$ (because for an arbitrary positive p the sum of all numbers $1/2^k$ less than p does not exceed $2p$: $\Sigma\{2^{-k} | 2^{-k} < p\} \leqslant 2p$). Note also that for each i the best approximation to p_i differs from p_i by no more than a factor 2. This trick reduces the problem to the first case.

More precisely, let us consider the enumerable set of pairs $\langle i, k\rangle$ such that $2^{-k} < p_i$. Its members form a computable sequence $\langle i(0), k(0)\rangle$, $\langle i(1), k(1)\rangle$, ...; the inequality $\sum_s 2^{-k(s)} < \infty$ holds since

$$\sum_s 2^{-k(s)} = \sum_i \Big(\sum_{i(s)=i} 2^{-k(s)}\Big) \leqslant \sum_i (2p_i) < \infty.$$

Therefore we can find a computable sequence of incomparable binary words $y(0)$, $y(1)$, ... such that $l(y(s)) \leqslant k(s) + c$ for some fixed c and for all s. Now we consider a computable mapping f such that $f(y(s)) = x_{i(s)}$. Then $KM_f(x_i) \leqslant k + c$ for all i, s such that $i(s) = i$, $k(s) = k$. Assume that i is fixed. Consider the least k such that $2^{-k} < p_i$. Then $p_i/2 \leqslant 2^{-k} < p_i$. There is an s such that $i(s) = i$, $k(s) = k$. For this s we have $k(s) \leqslant -\log_2 p_i + 1$, hence $KM_f(x_i) \leqslant -\log_2 p_i + 1 + c$. Q.E.D.

Remark. The inequality (2) of the theorem proved in this section can be strengthened:

$$KM\ (x) \leqslant -\log_2 m\ (x) + O\ (\log_2 (-\log_2 m\ (x))).$$

§5.4. A priori probability and randomness

In the previous section we saw that monotone entropy is close to the logarithm of a priori probability. So we can expect that both measures can be used for characterization of randomness. The first measure was discussed in Ch. IV; here we give a characterization of randomness (that is, typicalness or chaoticness—these properties are equivalent) in terms of the a priori probability.

Theorem. *Let P be a computable probability distribution on the space Ω of all infinite sequences of zeros and ones. Then for any sequence $\omega \in \Omega$ the following properties are equivalent:*

(1) ω is typical (= chaotic) with respect to the distribution P;

(2) the difference $-\log_2 P(\Omega_{(\omega)_n}) - (-\log_2 m((\omega)_n))$ is bounded [we recall that $(\omega)_n$ is the prefix of ω having length n].

This theorem is a variant of the Levin–Schnorr theorem obtained by replacing $KM((\omega)_n)$ by $-\log_2 m((\omega)_n)$. The condition (2) can be reformulated as follows: the quotient $P(\Omega_x)/M(\Sigma_x)$ is bounded and separated from zero for all x that are initial segments of ω. (We write $P(\Omega_x)$ but $M(\Sigma_x)$ because the probability distribution P is defined on the set Ω of infinite sequences and the distribution M is defined on the set Σ of finite and infinite sequences.) Let us note that the quotient $P(\Omega_x)/M(\Sigma_x)$ is bounded because the measure M is maximal (in the case of monotone entropy the inequality $KM(x) \leqslant -\log_2 P(\Omega_x) + O(1)$ can be shown to be true for all x for similar reasons). So we have to prove that ω is typical if and only if $P(\Omega_{(\omega)_n})/M(\Sigma_{(\omega)_n})$ is separated from zero.

Proof. 1. Let us assume that $P(\Omega_{(\omega)_n})/M(\Sigma_{(\omega)_n})$ is not separated from zero and prove that ω is not typical. For each rational number $\varepsilon > 0$ we consider the set A_ε of all words x such that $P(\Omega_x)/M(\Sigma_x) < \varepsilon$. This set is enumerable because the distribution P is computable and the a priori probability m is semicomputable from below. It is easy to see that the set $S_\varepsilon = \bigcup\{\Omega_x | x \in A_\varepsilon\}$ has a small measure with respect to the probability distribution P, namely, $P(S_\varepsilon) \leqslant \varepsilon$. Indeed, let $x_0, x_1, \ldots$ be all the elements of A_ε that are not continuations of other elements of A_ε. Then $S_\varepsilon = \bigcup \Omega_{x_i}$ and

$$P(S_\varepsilon) = \sum P(\Omega_{x_i}) \leqslant \varepsilon \cdot \sum M(\Sigma_{x_i}) \leqslant \varepsilon \cdot 1 = \varepsilon,$$

because the sets Σ_{x_i} are disjoint. If the quotient $P(\Omega_{(\omega)_n})/M(\Sigma_{(\omega)_n})$ is not separated from zero, then the sequence ω belongs to S_ε for all ε and, therefore, is not typical. (Here is it necessary to use Lemma 3 from §4.3 in the same way as in §4.1.)

2. Let us prove the reverse implication. We can see that it is a direct consequence of the Levin–Schnorr theorem. Indeed, if the sequence ω is not typical, then (according to the Levin–Schnorr theorem) the difference $-\log_2 P(\Omega_{(\omega)_n}) - KM((\omega)_n)$ is not bounded from above and $-\log_2 m((\omega)_n) \leqslant$ $\leqslant KM((\omega)_n) + O(1)$. Nevertheless we give a direct proof.

Assume that ω is a non-typical sequence. For each $\varepsilon > 0$ there is a computable sequence of binary words $X(\varepsilon, 0), X(\varepsilon, 1), \ldots$ such that $\omega \in \bigcup_i \Omega_{X(\varepsilon,i)}$ and $\Sigma P(\Omega_{x(\varepsilon,i)}) < \varepsilon$. Let us consider a measure that is zero outside $\bigcup_i \Omega_{X(\varepsilon,i)}$ and is equal to P/ε inside. More precisely, let us consider a measure P_ε on the set Σ such that

$$P_\varepsilon(\Sigma_u) = (1/\varepsilon)\cdot P\,(\Omega_u \cap \bigcup_i \Omega_{X(\varepsilon,i)}).$$

(Strictly speaking, this equality does not define a probability distribution because the measure of the whole space Σ defined by means of it is not equal to 1 but only less than or equal to 1. We improve P_ε by assuming that $P_\varepsilon(\Sigma) = 1$; then P_ε is a probability distribution on $\Omega \cup \{\Lambda\}$.) Now we take a sequence ε_s of positive rational numbers which converges to zero fast enough and a converging series Σk_s with terms converging to zero not too fast. It is enough to take $\varepsilon_s = 1/2^s$ and $k_s = 1/s^2$. Now we consider a probability distribution $Q = c\Sigma k_s P_{\varepsilon_s}$, where the constant c is chosen in such a way that the measure of the whole space is equal to 1. We shall prove that for our non-typical sequence ω the quotient $P(\Omega_{(\omega)_n})/Q(\Sigma_{(\omega)_n})$ and, therefore, the quotients $P(\Omega_{(\omega)_n})/M(\Sigma_{(\omega)_n})$ are not separated from zero. Indeed, for each s the sequence ω has an initial segment equal to $X(\varepsilon_s, i)$ for some i (otherwise $\omega \notin \bigcup \Omega_{X(\varepsilon_s,i)}$, which contradicts our assumption). Let us denote this initial segment by x. Evidently, $Q\,(\Sigma_x) \geqslant c\cdot k_s\cdot P_{\varepsilon_s}\,(\Sigma_x) = c\cdot k_s\cdot(1/\varepsilon_s)\,P\,(\Omega_x)$. So $P(\Omega_x)/Q(\Sigma_x) \leqslant \varepsilon_s/ck_s$. But $\varepsilon_s/ck_s \to 0$ as s tends to infinity. The theorem is proved.

Actually, we have simply repeated the proof of the Levin–Schnorr theorem with some simplifications connected with the replacement of the monotone entropy by the a priori probability.

Let us point out that our proof establishes more than we claimed: we have proved in fact that if ω is not typical, then $P(\Omega_{(\omega)_n})/M(\Sigma_{(\omega)_n})$ converges to zero. (Indeed, in the inequality at the end of proof we can replace the word x by its arbitrary continuation.) So we may say "does not tend to infinity" instead of "is bounded" in the statement of the theorem. (Compare the similar remark at the end of §4.1 concerning the Levin–Schnorr theorem.)

CHAPTER VI

THE FREQUENCY APPROACH TO THE DEFINITION OF A RANDOM SEQUENCE

§6.1. Von Mises' approach. The Church and Kolmogorov–Loveland definitions

The frequency approach was suggested by von Mises (a German mathematician and mechanician) in [41] and [42]. We must realize, however, that von Mises considers the notion of a random sequence ("Kollektiv" in his terms) as a fundamental notion of probability theory. From our modern viewpoint this notion is defined in the context of measure-theoretic probability theory: the notion of a probability distribution is a primary notion, and then we define the notion of a random object with respect to this distribution.

On the contrary, von Mises considered the notion of the "Kollektiv" as the primary one and the notion of a probability distribution as an attribute of a "Kollektiv". We must bear in mind also that von Mises' approach was not a mathematical one according to modern set-theoretic standards. For example, von Mises used the notion of a "legal selection rule" without formal definition and considered the existence of gambling houses as an argument showing that "Kollektives" do exist.

After these warnings we shall try to present von Mises' viewpoint. Let us consider a sequence of zeros and ones obtained by tossing a symmetric coin (zero denotes head, one denotes tail). It is known from our experience that the fraction of ones in the initial segment of the sequence having length N tends to 1/2 as N tends to infinity. This fact can be interpreted as follows: the games with the symmetric coin (when the head appears we win a cent, otherwise we lose a cent) is fair (on the average we win nothing and lose nothing).

Moreover, the experience of gambling houses shows that no "gambling system" (saying when we make a bet and when we pass) can guarantee a systematic gain. This "gambling system" can be regarded as a "selection rule", which selects a subsequence of a sequence obtained by coin tossing (selected terms correspond to the coin tossings when a bet is made). For a random sequence the limit of the frequency of ones in a subsequence selected by such a rule is equal to 1/2 (as for the whole sequence). Of course, some selection rules ("gambling systems") are not admissible. Here is an evident example of a non-admissible rule: "select terms equal to 1" ("make a bet only if you win").

Similar properties hold for an asymmetric coin. In this case the frequency of ones in the whole sequence tends to some p $(0 < p < 1)$ and the same is true for any subsequence selected by an admissible selection rule. The number p is called the *probability* of tails.

So we can describe the idea of von Mises as follows. We consider so-called selection rules. Any selection rule can be applied to an infinite sequence of zeros and ones and gives a (finite or infinite) subsequence of it. Some selection rules are *admissible*. Each infinite subsequence obtained by the application of an admissible selection rule to a given sequence is called a *legal* subsequence of it (the sequence itself also consititutes a legal subsequence). An infinite sequence of zeros and ones is random (in the sense of von Mises) with respect to the Bernoulli distribution with probability of ones equal to p if the following property holds: for any legal subsequence the limit of frequencies of ones in its initial segments exists and is equal to p.

Thus we have given a brief description of von Mises' original idea. Later Church proposed the following formal definition of an admissible seletion rule. First of all the decision (to make a bet or to pass) must be made on the basis of the result of previous coin tossings. In other words, an admissible rule of choice is a set A of binary words. Applied to a sequence $x_0x_1...$ it selects terms x_n such that the word $x_0x_1 \dots x_{n-1}$ belongs to A (these terms go in the same order as in the given sequence).

But this requirement is not enough, because for each sequence $x_0x_1\dots$ we can consider the set A containing all words $x_0\dots x_{n-1}$ such that $x_n = 1$. The corresponding selection rule evidently gives a subsequence containing only ones. In this case von Mises would probably say that this selection rule is not admissible, because the selection rule must be fixed before the game. But it is not clear how to express ths requirement mathematically, and Church replaced it by the requirement of the decidability of the set A. (A is called *decidable* if there is an algorithm that can decide whether an arbitrary given word x belongs to A or not.)

Later Kolmogorov [17] and independently Loveland ([36], p. 499) generalized the notion of an admissible selection rule. This generalization is connected with a different scheme for a game: now the player may select the order of observations of terms of a given sequence. Imagine that zeros and ones forming the sequence are written on cards presented to the player in such a way that he sees only their blank back sides. The player is allowed to turn over any card (not yet turned over) without a bet. He may also bet on a card not turned over yet. In this case he wins a cent if 1 is written on the card and loses a cent if 0 is written. This scheme corresponds to a rule selecting cards on which a bet was made (in the order they were turned over). Let us mention that now the notion of a subsequence is more general than usual: the order of terms in the subsequence may differ from their order in a given sequence.

Let us give a formal definition. An admissible selection rule (in the sense of Kolmogorov and Loveland) is a pair of functions F and G. Both F and G have binary words as arguments; the values of F are natural numbers, the value of G are Boolean values True and False. (The function F says which card must be turned over next, the function G says whether a bet is made.) Let us mention that the functions F and G may be partial. Now we describe the application of this selection rule to a sequence $x_0x_1\dots$. First we define a sequence of natural numbers $n_0 = F(\Lambda)$ (Λ is the empty binary word), $n_1 = F(x_{n_0})$, $n_2 = F(x_{n_0}x_{n_1})$, and so on; the construction terminates if at least one of the values $F(x_{n_0}x_{n_1} \dots x_{n_k})$ and $G(x_{n_0}x_{n_1} \dots x_{n_k})$ turns out to be undefined or the value $F(x_{n_0}x_{n_1} \dots x_{n_k})$ coincides with one of the values $n_0, n_1, \dots, n_k$. Then we select among the n_k those for which $G(x_{n_0}x_{n_1} \dots x_{n_{k-1}})$ is true; the corresponding x_{n_k} constitute the selected subsequence (in the order of increase of k).

Let p be an arbitrary (not necessarily computable) real number from $(0, 1)$. A sequence ω is called *Church stochastic* (or *von Mises–Church random*) with respect to the Bernoulli distribution with probability of ones equal to p if the frequency of ones in its initial segments tends to p and the same is true for all infinite subsequences obtained by a Church admissible selection rule.

Replacing "Church admissible" by "Kolmogorov–Loveland admissible" we obtain the definition of a *Kolmogorov–Loveland stochastic* (or *von Mises–Kolmogorov–Loveland random*) sequence.

The properties of these definitions are discussed in the next sections of this chapter. Now let us conclude this section by a simple remark regarding the case of a symmetric coin. Let us give more freedom to a player. Allow him to make bets both on zeros and ones. This means that before a coin tossing (in the Church scheme) or before turning over a card (in the Kolmogorov–Loveland scheme) a player may say "I bet on zero", "I bet on one", or "pass". In the third case he wins nothing and loses nothing. In the first two cases he wins one cent if he guessed correctly or loses one cent otherwise. The sequence is called *stochastic* if the average gain of the player (gain divided by the number of bets) tends to zero for an arbitrary game system.

It is easy to see that this new freedom does not change the class of stochastic (Church stochastic or Kolmogorov–Loveland stochastic) sequences. Indeed, assume that we have a "gambling system" S in the new sense (we are allowed to put bets on zeros or ones). Consider two "gambling systems" S_0 and S_1 in the old sense. The system S_0 is obtained if we make bets when S bets on zero, and the system S_1 is obtained if we make bets when S bets on one. It is clear that the gain of the system S is equal to the difference (the gain of S_1)—(the gain of S_0). Therefore the class of stochastic sequences remains the same.

Question. Apply some Kolmogorov–Loveland admissible rule to a Kolmogorov–Loveland stochastic sequence. Is the selected subsequence necessarily Kolmogorov–Loveland stochastic? (This question is discussed in [50].)

§6.2. Relations between different definitions. Ville's construction. Muchnik's theorem. Lambalgen's example

6.2.1. Relations between different definitions.

We now have definitions of different versions of the notion of a stochastic sequence. It is natural to compare them with the definition of a typical sequence (equivalent to the definition of a chaotic sequence). This comparison can be made only for Bernoulli distributions (otherwise stochasticness is undefined) with a computable probability of ones (otherwise typicalness and chaoticness are undefined). In this case the following theorem holds.

Theorem.

(a) *every typical sequence is Kolmogorov–Loveland stochastic*;

(b) *every Kolmogorov–Loveland stochastic sequence is Church stochastic*;

(c) *the reverse implications for* (a) *and* (b) *are false.*

In this section we give schemes of proofs for the assertions (a) – (c) and for some related results and constructions interesting in their own right. For simplicity we restrict ourselves to the case of the uniform Bernoulli distribution.

The assertion (b) is an immediate consequence of the definitions. To prove assertion (a) we recall the discussion of the law of lage numbers in §2.1. There we fixed a number $\varepsilon > 0$ and considered the sets D_n containing all sequences of zeros and ones such that the frequency of ones in their initial segment having length n differs from 1/2 by more than ε. The standard estimate (using Stirling's formula, the de Moivre–Laplace theorem, and so on) shows that the (uniform Bernoulli) measure of the set D_n exponentially decreases as n tends to infinity (we recall that ε is fixed). Each D_n can be represented as a finite union of disjoint intervals (corresponding to words having length n and frequency of units differing from 1/2 by more than ε). Therefore, the set $E_k = \bigcup\limits_{n \geqslant k} D_n$ can be represented as a union of a computable sequence of intervals; the sum of their measures can be made arbitrarily small (when k is large enough). So the set $\bigcap\limits_k E_k$ is an effectively null set and any sequence contained in it is not typical. Each sequence having no limit of frequencies or having limit of frequencies not equal to 1/2 belongs to $\bigcap\limits_k E_k$ for some ε, thus all typical sequences have the limit of frequences equal to 1/2.

We must prove also that for each Kolmogorov–Loveland admissible selection rule R the subsequence obtained by applying R to a typical sequence is finite or has a limit of frequences equal to 1/2. Let us assume that ε is fixed. Consider the set D_n^R containing all sequences ω such that application of the rule R to ω gives a sequence of length at least n and the frequency of ones in the first n terms of this sequence differs from 1/2 by more than ε. The measure of the set D_n^R does not exceed the measure of the set D_n (but can be less, since the application of R can give a finite sequence having length less than n). The set D_n^R can be represented as the union of an enumerable set of intervals (if $\omega \in D_n^R$, then this is guaranteed by a finite initial segment of the sequence ω). After these remarks the proof goes as before.

Remark. It is essential that we use Martin-Löf's definition of a typical sequence but not Schnorr's, since we cannot guarantee the effective convergence of the series of measures of intervals forming D. Not all Schnorr typical sequences are Kolmogorov–Loveland stochastic; any Schnorr typical sequence with logarithmically increasing entropies of initial segments (see §2.3.1) is not Kolmogorov–Loveland stochastic because of Muchnik's theorem (see below).

6.2.2. Ville's example.

We begin the proof of assertion (c) with a construction due essentially to J. Ville. He showed that there is a Church stochastic sequence such that any initial segment of it has the following property: the number of zeros in it is not less than the number of ones. (Actually Ville's construction does not depend on the decidability requirement mentioned in Church's definition of an admissible selection rule, and was invented before Church's definition. To explain it we follow [35].) The set of all sequences possessing the above mentioned property is an effectively null set. (The so-called law of the iterated logarithm implies that it is a null set; the analysis of the proof of the law shows that this null set is in fact an effectively null one.) So the sequence constructed by Ville is not a typical one. Thus Ville's example shows that there is a Church stochastic sequence that is not typical, therefore at least one of the inverse implications for the assertions (a) and (b) of the theorem is false.

Now we shall explain Ville's construction. We want to construct a sequence such that the application of any Church admissible selection rule gives a "balanced" subsequence (we call a sequence *balanced* if the frequency of ones in its initial segments tends to 1/2). Let us begin with a model example and consider only one admissible selection rule.

We shall construct a sequence by induction. Assume that we have constructed an initial segment $x_0x_1 \dots x_k$. Looking at the selection rule we can find out whether the next term x_{k+1} will be included in the selected subsequence. If it is included, let q be its number in the selected subsequence (it is 1 for the first selected term, and so on). If the term is not included, let q be its number in the subsequence formed by terms not selected. The value of x_{k+1} is equal to $(q+1)$ mod 2. So the selected subsequence will be 01010101...; the same is true for the sequence formed by terms not selected. Each initial segment of the sequence $x_0x_1x_2 \dots$ is a "mixture" of two initial segments of the sequence 01010101... and, therefore, the number of zeros in it is not less than the number of ones.

Assume that we have m selection rules $R_1, \dots, R_m$. Then each term of a sequence can be characterized by an m-bit vector describing which rules (among $R_1, \dots, R_m$) select this term. So our sequence is a mixture of 2^m subsequences. Each subsequence corresponds to a specific value of this binary vector. We will construct the sequence in such a way that each of these 2^m subsequence has the form 01010101... . This requirement determines the sequence uniquely and guarantees that for any initial segment of the sequence the number of zeros in it is not less than the number of ones. Applying the i-th selection rule R, we obtain a subsequence which is a "mixture" of 2^{m-1} subsequences having the form 01010101... (corresponding to 2^{m-1} bit vectors having 1 in the i-th place). Evidently, this mixture is balanced.

Now let us consider the case when there are countably many selection rules. In this case each term of the sequence is characterized not by a bit vector but rather by an infinite sequence of bits $u_0u_1 \dots$ ($u_i = 1$ if the i-th rule selects this term, $u_i = 0$ otherwise). In other words, each term is characterized by an infinite path in a binary tree. Nevertheless, we shall use only an initial segment of this path. More precisely, let us choose a sequence of natural numbers $n_0 < n_1 < \ldots$ growing fast enough, for example, $n_i = 2^{2i}$. At each stage of our construction (that is, for each term of the sequence) one of the vertices of the binary tree will be called active. To find the active vertex we start from the root and follow the path (corresponding to the already constructed initial segment of the sequence) until we find a vertex which was active less than n_i times, where i is its height. In other words, we choose as active the shortest binary word x such that

(1) the m-th digit in x is equal to 1 if and only if the term considered is selected by the m-th rule R_m;

(2) the word x was active up to now less than n_i times.

So the sequence is divided into countably many subsequences. For any binary word x of length i (that is, for any vertex in the binary tree having height i) the corresponding sequence has length at most n_i. It contains all terms corresponding to those stages of the construction for which x was active. The sequence corresponding to a binary word (vertex) x starts only when all sequences corresponding to initial segments of x are finished.

It remains to describe how the values of the terms of the considered sequence are chosen. They are chosen in order to satisfy the following condition: for each binary word x the sequence corresponding to x has the form 01010101... . This requirement can be fulfilled, since every time before choosing the value of the next term we know already which subsequence it belongs to. This requirement guarantees that for each initial segment the number of zeros in it is not less than the number of ones. It remains to prove that any selection rule R_i selects a balanced subsequence.

Let $y_0y_1 \ldots$ be an infinite subsequence obtained by the application of the i-th selection rule R_i. Let us consider an arbitrary initial segment of the sequence $y_0y_1 \ldots$ and the binary words (vertices of the binary tree) corresponding to the terms of the segment. The terms can be divided into two groups. For some of them the corresponding words have length at most i; the total number of such terms is bounded (it is not greater than $2^0n_0 + \ldots + 2^in_i$), so we may ignore them. For other terms the corresponding words have length greater than i and their i-th bit is equal to 1 (because they are selected by the i-th rule). Let x be the longest word among words corresponding to some term in the initial segment $y_0y_1 \ldots$ considered. Let k be its length. (As we have already said, we may assume that $k > i$.) So the total number of the words used does not exceed $1+2+ \ldots +2^k < 2^{k+1}$. The numbers of zeros in the sequence corresponding to each word is equal to the number of ones or exceeds the latter by 1. Thus the difference between zeros and ones in the initial segment of $y_0y_1 \ldots$ considered does not exceed 2^{k+1}. Let us show that the length of this initial segment is large enough. Indeed, if the word x is used as active, then the word $x' =$ (x without the last bit) is used (as active) completely, that is, n_{k-1} times. The i-th bit of x' is equal to 1 (we assume that $k > i$), therefore, all the terms corresponding to x' are selected by R_i. So the length of the initial segment considered is at least $n_{k-1} = 2^{2k-2}$, and the frequency of ones is close to 1/2 (the difference is less than $(2^{k+1})/(2^{2k-2})$ and tends to zero).

We have constructed a Church stochastic sequence such that for any initial segment the number of zeros in it is not less than the number of ones. (To obtain a Church stochastic sequence such that the number of zeros in its initial segments is greater than the number of ones it is enough to add a leading 0.) This construction does not use the algorithmic nature of the selection rule; any countable family of selection rules can be used. It is important that all rules consider the terms of a sequence in the same order. So this construction cannot be extended to Kolmogorov–Loveland admissible rules.

Looking at Ville's construction more closely, we can establish the existence of Church stochastic sequences for which the entropy of the initial segment of length n is $O(\log n)$. Indeed, this construction enables us to construct for each countable family R_0, R_1, ... of selection rules a sequence ω balanced with respect to all R_i. If we have an algorithm enumerating all these rules (giving a program for R_i from a given i) the sequence ω will be computable. If an algorithm enumerating all Church admissible rules existed, then we would obtain a computable Church random sequence (this is, of course, impossible). So the construction of a Church random sequence cannot be effective and requires additional information (saying which programs correspond to Church admissible rules, that is, decidable sets). But the amount of necessary information can be small in comparison with the length of the initial segment if the numbers n_i used in the construction grow fast enough. This enables us to construct a Church stochastic sequence with logarithmically increasing entropies of the initial segments. So we have another example of a Church stochastic sequence that is not typical and chaotic.

6.2.3. Muchnik's theorem.

The existence of a Church stochastic sequence with logarithmically increasing entropies of initial segments was mentioned (without proof) in [19]. In the same paper Kolmogorov claims that there exist Kolmogorov–Loveland stochastic sequences with logarithmically increasing entropies of initial segments. This assertion is false, as Andrei A. Muchnik recently showed. Namely, he proved the following theorem.

Theorem (Muchnik). *Let ω be an infinite sequence of zeros and ones and suppose that the entropy of an initial segment of ω having length n does not exceed αn for some $\alpha < 1$ and for all sufficiently large n. Then ω is not Kolmogorov–Loveland stochastic.*

This theorem is as yet unpublished. Muchnik kindly permitted us to reproduce here the sketch of the proof. Let n be a natural number and A a set of n-bit binary words. We consider the following game between a player and his opponent. The opponent chooses from the set A an arbitrary sequence of zeros and ones and writes its terms on cards. The player sees only the blank backs of cards. The player turns the cards over in the same order as in the sequence of zeros and ones. Before turning each card over he can put an arbitrary amount of money not exceeding 1 cent on zero or one. If he guesses correctly, he wins this amount of money, otherwise he loses the same amount of money. (This scheme differs from the Church scheme in two respects: 1) a finite number of cards; 2) an arbitrary real gain between -1 and $+1$ instead of three possibilities -1, 0, $+1$.)

***Lemma* 1.** *Assume that $\alpha < 1$. Then for each natural number n and for each set $A \subset \{0, 1\}^n$ with cardinality at most $2^{\alpha n}$ we can find a strategy for the player which guarantees him a gain of at least $(1-\alpha)n$ cents.*

Proof of Lemma 1. Let us define the notion of *information capital* of the player (at a given stage of the game). Let $x_0 \dots x_{k-1}$ be the values on the cards already turned over. Let M be the total number of continuations of the sequence $x_0 \dots x_{k-1}$ having length n ($M = 2^{n-k}$), and m the number of continuations of the sequence $x_0 \dots x_{k-1}$ that are elements of A. We define the information capital as $\log_2(M/m)$. At the beginning of the game the information capital is not less than $(1-\alpha)n$, at the end of the game it is equal to zero. Let us show that there is a strategy for the player which guarantees that the sum of his gain and his information capital does not decrease during the game. (If this is so, at the end of the game the information capital is equal to 0 and the gain is not less than $(1-\alpha)n$.)

To prove the existence of such a strategy we consider (for each stage of the game) the numbers

m_0—the number of continuations of the sequence $x_0 \dots x_{k-1}0$ that are elements of A;

m_1—the number of continuations of the sequence $x_0 \dots x_{k-1}1$ that are elements of A.

(Evidently, $m = m_0+m_1$.) The quotients $p_0 = m_0/m$ and $p_1 = m_1/m$ can be regarded as the conditional probabilities of zero and one after $x_0 \dots x_{k-1}$ (if all elements of A are regarded as equiprobable a priori); evidently, $p_0+p_1 = 1$. We must choose a stake between -1 and $+1$ (negative values correspond to bets put on one, positive values correspond to bets put on zero) in such a way that in all cases the sum of the gain of the player and his information capital does not decrease. Let x be the amount of money put on zero. If zero appears, then $(-\log_2 p_0-1)$ is added to the information capital and $(x-\log_2 p_0-1)$ is added to the sum of the gain and the information capital. If one appears, then $(-x-\log_2(1-p_0)-1)$ is added to the sum of the gain and the information capital. So it remains to prove that for each $p \in [0, 1]$ there is an $x \in [-1, 1]$ such that both numbers $(x-\log_2 p-1)$ and $(-x-\log_2(1-p)-1)$ are non-negative. Let a and b be arbitrary real numbers. The existence of $x \in [-1, 1]$ such that $x-a \geqslant 0$ and $-x-b \geqslant 0$ is equivalent to the conjunction of the following two conditions: 1) $a \leqslant -b$; 2) $[a, -b] \cap [-1, 1] \neq \varnothing$. So in our case it is enough to prove that $\log_2 p+1 \leqslant -\log_2(1-p)-1$, that is, $(\log_2 p+\log_2(1-p))/2 \leqslant -1$ (this is a consequence of the convexity of the logarithm function) and that at least one of numbers $\log_2 p+1$ and $-\log_2(1-p)-1$ belongs to $[-1, 1]$ (the first if $p \geqslant 1/2$, the second if $p \leqslant 1/2$). Lemma 1 is proved.

The next lemma deals with strategies that always make maximal stakes (one cent on zero or one). To compensate for this restriction we shall consider a set of strategies instead of one strategy.

***Lemma* 2.** *Let $\alpha < 1$. Then for each n and for each $A \subset \{0, 1\}^n$ such that the cardinality of A does not exceed $2^{\alpha n}$ we can find a finite set of strategies $S_1, \ldots, S_t$ making maximal stakes such that for each sequence $x \in A$ there is a strategy S_i with gain $\geqslant ((1-\alpha)/2)n$ on the sequence x. The number of strategies depends only on α (but not on n and A).*

Proof of Lemma 2. Let us consider a stragtegy S with arbitrary stakes, which exists by Lemma 1. Let us consider a strategy S' with stakes that are multiples of $1/N$ for some natural number N; S' is an approximation to S in the sense that in any case the difference between the stakes made by S and S' does not exceed $1/N$. If N is large enough $(1/N < (1-\alpha)/2)$, then S' guarantees the gain $((1-\alpha)/2)n$. Now let us represent S' as the arithmetic mean of $2N$ strategies $S_1, \ldots S_{2N}$ with maximal stakes. (For example, assume that S puts m/N on zero. In this case some of $S_1, \ldots, S_{2N}$ will put a stake on 0 and others on 1; there will be $N+m$ strategies of the first type and $N-m$ strategies of the second type.) For each sequence $x \in A$ the following statement is true: the gain of S' on x is the arithmetic mean of the gains of the strategies $S_1, \ldots S_{2N}$. So at least one of the strategies $S_1, \ldots, S_{2N}$ must have a gain of at least $((1-\alpha)/2)n$. Lemma 2 is proved.

Now we proceed to the proof of Muchnik's theorem. Let ω be a sequence such that the entropy of its initial segment of length n does not exceed αn (for some $\alpha < 1$ and for all sufficiently large n). We cut the sequence ω into pieces $u_0, u_1, \ldots$ of length $n_0, n_1, \ldots$ $(\omega = u_0 u_1 \ldots)$. If the numbers n_i increase fast enough (for example, $n_i = 2^{2^i}$) then the entropy of u_i is less than $\lfloor \beta \cdot l(u_i) \rfloor$ for some rational constant $\beta < 1$ and for all sufficiently large i (for simplicity we assume that this holds for all i ignoring a finite initial segment); $\lfloor z \rfloor$ stands for the integer part of z. So the word u_i belongs to the set A_i of all words having length n_i and entropy less than $\lfloor \beta n_i \rfloor$; the number of elements in this set does not exceed $2^{\beta n_i}$. Using Lemma 2 we can find for each i a set of strategies $S_1, \ldots, S_t$ such that for each element $u \in A_i$ (therefore, for u_i too) at least one of the strategies S_j has a gain of at least $((1-\beta)/2)n_i$. The number of strategies (t) does not depend on i, so we can form t strategies for playing with the whole sequence. One of these t strategies infinitely often will have gain greater than $\varepsilon \cdot$ (number of bets) for some $\varepsilon > 0$.

To prove this it is sufficient to note that the gain in the game with u_i is substantially greater than any possible loss before, and that there is a strategy that is successful infinitely many times (because each time one of the strategies is successful).

This argument would show that the sequence ω is not Church stochastic if the strategy we have constructed were computable. (We recall that the right to make bets on zeros and ones was discussed at the end of §6.1.) But this is not so, becuase the list of all words having length n_i and entropy less than $\lfloor \beta n_i \rfloor$ cannot be computed effectively from a given i; we can enumerate them (if a word has a short description, this fact will become known) but we cannot be sure that all such words have already appeared.

Muchnik overcomes this difficulty as follows. We group the segments u_0, u_1, ... into pairs u_0u_1, u_2u_3, ... We know that the word u_{2n} belongs to A_{2n} and the word u_{2n+1} belongs to A_{2n+1}. We have two possibilities (in the Kolmogorov–Loveland scheme) to play with the sequence $u_{2n}u_{2n+1}$. First possibility: we turn the cards over and learn the word u_{2n}. We then enumerate A_{2n} until u_{2n} is found. Then we make the same number of steps enumerating A_{2n+1}; the part of A_{2n+1} found during this process is denoted by $\bar{A}_{2n+1}$ and used instead of A_{2n+1} when we construct a strategy playing with u_{2n+1}. The other possibility is symmetrical. We turn the cards over and learn u_{2n+1},. enumerate A_{2n+1} until u_{2n+1} is found, make the same number of steps enumerating A_{2n}, denote the discovered part of A_{2n} as $\bar{A}_{2n}$ and use it in a strategy playing with u_{2n}. We may be sure that at least one of these two strategies will be successful (if u_{2n} appears in the enumeration of A_{2n} before u_{2n+1} appears in A_{2n+1}, then the use of $\bar{A}_{2n}$ instead of A_{2n} is legal and the second strategy is successful; otherwise the first one is successful).

For each segment $u_{2n}u_{2n+1}$ we have t pairs of strategies, so we have $2t$ computable strategies in the Kolmogorov–Loveland game with the infinite sequence. We denote them by $S_{pr}(1 \leqslant p \leqslant t,\ r = 0 \text{ or } 1)$. The strategy S_{p0} learns (without bets) u_0, u_2, u_4, ... and uses the p-th strategy based on the sets $\bar{A}_1$, $\bar{A}_3$, $\bar{A}_5$, ...for the segments u_1, u_3, u_5 ...; the strategy S_{p1} learns (without bets) u_1, u_3, u_5, ... and uses the p-th strategy based on the sets $\bar{A}_0$, $\bar{A}_2$, $\bar{A}_4$, ... for the segments u_0, u_2, u_4 ... . Now all strategies are computable. For each n either u_{2n} appears in the enumeration of A_{2n} earlier than u_{2n+1} appears in the enumeration of A_{2n+1} or vice versa. In the first case one of the strategies S_{p1} will be successful, in the second case one of the strategies S_{p0} will be successful. We have a finite number of strategies, hence at least one of them will be successful infinitely many times and, therefore, the sequence ω is not Kolmogorov–Loveland stochastic. Muchnik's theorem is proved.

6.2.4. Lambalgen's example.

Let us show that the inverse implication for the assertion (a) of Theorem 6.2.1 is false. This can be done by using the method of Lambalgen [24], [25]. In this paper he gave a new proof of the existence of a Church stochastic sequence that is not typical. As we remarked in [54], the same method enables us to construct a Kolmogorov–Loveland stochastic sequence that is not typical. We do not go into the details of this construction, but its idea is simple. Let us consider a computable sequence of rational numbers p_0, p_1, ... converging computably to 1/2. Let us consider the probability distribution of the results of independent trials such that the probability of a success in the n-th trial is p_n. We denote this probability distribution on Ω by μ. We can prove that in this case each sequence typical with respect to μ will be Kolmogorov–Loveland stochastic with respect to the uniform Bernoulli distribution. However, if the p_i converge to 1/2 slowly ($\Sigma(p_i - 1/2)^2 = +\infty$), then no sequence typical with respect to μ will be typical with respect to the uniform Bernoulli distribution. It remains to use, for example, $p_i = (i+10)^{-1/2} + 1/2$. (See the details in [54].)

It remains to show that the inverse implication for the assertion (b) of the theorem is false (that is, there are sequences that are Church stochastic but not Kolmogorov–Loveland stochastic). An example of such a sequence was consructed by Loveland [35]. He constructed a Church stochastic sequence which becomes not Church stochastic after a computable permutation of its terms. Evidently, this is impossible for a Kolmogorov–Loveland stochastic sequence.

We do not reproduce his construction here, because Muchnik's theorem implies that a Church stochastic sequence with the logarithmically increasing entropies of initial segments mentioned above cannot be Kolmogorov–Loveland stochastic.

§6.3. A game-theoretic criterion for typicalness

In this section we show that the criterion for typicalness given in §5.4 can be (for the case of the uniform Bernoulli distribution) reformulated in game-theoretic terms. The corresponding game can be regarded as a generalization of the games described above. We consider a game where the stakes may be arbitrary real numbers between 0 and 1. Let us change the rules and require that the size of the stakes is restricted by the current capital of the player. More precisely, the player can divide all his capital into three parts. He bets the first part on zero, bets the second part on one, and throws out the third part (the last action seems evidently non-profitable, but it is necessary to allow it for reasons which will become clear later). The part bet on the correctly guessed digit is doubled, the part bet on the incorrectly guessed digit is lost (like the third part). So, dividing his capital into two equal parts and betting them on zero and one, the player in all cases wins or loses nothing. The initial capital is equal to 1. The terms of a sequence become known in their usual order (as in the Church scheme).

The strategy in the game described above is a rule which tells the player how he must divide his capital into three parts depending on the already known terms of the sequence. This strategy is uniquely determined by a function $L : x \mapsto L(x)$, where $L(x)$ is the capital of the player who acts according to this strategy playing with x. In terms of this function L the strategy may be described as follows: the capital (equal to $L(x)$) is divided into three parts: $L(x1)/2$, $L(x0)/2$, and $L(x)-L(x1)/2-L(x0)/2$. (The first part is bet on zero, the second part is bet on one, and the third part is thrown out.)

The function L must satisfy the following evident requirements: $L(x) \geqslant 0$ for all x, $L(\Lambda) = 1$, $L(x0)+L(x1) \leqslant 2L(x)$. Functions satisfying these requirements are in one-to-one correspondence with the strategies in the game described above.

On the other hand, such functions are in one-to-one correspondence with semimeasures (in the sense of §5.1): the semimeasure $x \mapsto L\,(x)/2^{l(x)}$ corresponds to the function L. Measures that are semicomputable from below correspond to functions that are semicomputable from below. So we may reformulate the typicalness criterion from §5.4 as follows. Let us call a strategy *semicomputable from below* if the corresponding function L is semicomputable from below. A sequence ω is typical if and only if there is no semicomputable (from below) strategy giving unbounded gain playing with ω. (Here "unbounded gain" can be replaced by "gain tending to infinity"; see the remark at the end of §5.4.)

The features of this game are now clear, because this game is just a trivial reformulation of the typicalness criterion in game-theoretic terms. This explains the presence of a third (thrown out) part of the capital (it corresponds to the positive measures of finite sequences with respect to a priori probability on Σ) and the unnatural requirement of semicomputability from below for the values of the function L (and, for example, not for the quoteients $L(x1)/L(x)$ and $L(x0)/L(x)$).

Addendum

A timid criticism regarding probability theory

We begin with an example from Pólya's book [44], Vol. II, Ch. XIV, part 7, p. 76. Assume that "...315672 attempts to cast five or six spots with a dice produced 106602 successes. If all dice cast were fair, ... we should expect about 315672/3 = 105224 successes ... . Thus, the observed number deviates from the expected number by ... 1378. Does such a deviation speak for or against the hypothesis of fair dice?"

This question is traditional and Pólya's answer is traditional too. "...our judgement depends on the solution of the following problem: Given that the probability of a success is 1/3 and that the trials are independent, find the probability that in 315672 trials the number of successes should be either more than 106601 or less than 103847" ([loc. cit.]; the last two numbers are the expected value 105524 plus and minus the deviation 1377, which is 1 less than the real deviation). It is easy to find that this probability is less than 2.10^{-7}, "...and so the underlying hypothesis of fair dice appears extremely unlikely" [loc. cit.].

As we see, the scheme of the argument is as follows. There is a set of possible outcomes (the set of all records of a series of 315672 trials in the case considered). There is a statistical hypothesis, that is, a probability distribution on the set of all outcomes (the hypothesis of a fair dice; according to it the trials are independent and the probabilities of all the numbers from 1 to 6 are equal; so all records are equiprobable). Lastly, there is an experimental result, that is, one of the possible outcomes (in our case a record with 106602 successes). We want to know whether it contradicts the statistical hypothesis. A procedure to do this is as follows. We choose an event that took place during the experiment (in our case the event "number of successes differs from the expected number by more than 1377"). We compute the probability of this event (in our case it is less than $2\cdot 10^{-7}$). If this probability is small, the statistical hypothesis is discredited. "The actual occurrence of an event to which a certain statistical hypothesis attributes a small probability is an argument against that hypothesis, and the smaller the probability, the stronger is the argument" [loc. cit.].

In our exposition we ignored the following difficulty: it is not clear what events one may consider. Why did we compute the probability of the event "deviation $\geqslant$ 1378" and not the probability of the event "deviation = 1378"? (In the latter case the probability is smaller.) We could also compute the probability of the event "the numbers on the dice are exactly the same as in the experiment" and this probability is very small. Using the latter event we can reject the hypothesis of a fair die independently of the result of the experiment.

We return to a question posed in the Introduction, where we considered the sequences of 12 zeros and ones as the records of a coin tossing. We said that the hypothesis of a fair coin is usually rejected if the record contains 12 zeros. The motives for this rejection are usually explained as follows: the probability of this event (12 zeros) is very small. But each sequence of 12 zeros and ones has the same probability!

Of course, this problem could not go unnoticed. Pólya comments [loc. cit.] on his example: "should we regard the deviation 1378 as small or large? Is the probability of such a deviation high or low? The last question seems to be the sensible question. Yet we still need a sensible interpretation of the short, but important, word "such". We shall reject the statistical hypothesis if the probability that we are about to compute turns out to be low. Yet the probability that the deviation should be exactly equal to 1378 units is very small anyhow– even the probability of a deviation exactly equal to 0 would be very small. Therefore, we have to take into account all the deviations of the same absolute value as, or of larger absolute value than, the observed deviation 1378".

We hope that you will agree with us that this argument does not seem convincing: why is one meaning of the word "such" better than others?! But the problem of a "reasonable meaning" is stated quite definitely.

Let us give some other quotations. Renyi writes in his "Letters on probability" [45], Russian ed., p. 153 (Pascal's imaginary letter): "In fact, what does the expression "the cards are well shuffled" mean? ...If the cards are well shuffled, then all orderings of them are equiprobable. But how can one say whether the cards are well shuffled by looking at their ordering if every two orderings have the same probability? And if it is impossible to decide whether cards are well shuffled by looking at their ordering, how can the expression "well shuffled" be meaningful?".

Objections of this kind have a long history. In the book "Le hasard" [5], Russian ed., p. 76 – 77, Emile Borel cited the following passages written by Bertrand (who invented a paradox showing that different methods of computing the probability of the event "a random chord contains more than 1/3 of the circumference" give different results):

"The Pleiades [a cluster of stars six of which are readily visible; the question is whether the stars form a cluster in space or their closeness on the sky is a casual coincidence—Authors' note] seem closer to each other than they should. This assertion seems reasonable, but if we try to express our opinion in figures our knowledge is not enough. How can we give a precise definition for this vague notion of "closeness"? Should we look for the smallest circle containing this group? The maximal angle distance? The sum of the squares of all distances? ... All these quantities are less than one can expect. Which of them can be used as a measure of a probability? If three stars form an equilateral triangle, should we conclude that this fact (which has a small probability a priori) must have a specific reason?"

Such objections have a long history, but there is no generally accepted answer to them. Borel writes in [5], Russian ed., pp. 77 – 78: "Let us comment on Bertrand's idea about the equilateral triangle formed by three stars; it is connected with the question of a round number. If we choose randomly a number between 1000000 and 2000000 then the probability that it is equal to 1342517 is equal to one millionth; the probability that it is equal to 1500000 is also one millionth. Nevertheless the second possibility is often considered as less probable; it is because nobody considers such a number as 1542317 individually; it is considered as a class of numbers of the same type; if we change one digit it is hardly noticed, and the number 1324519 does not differ from 1324517; a special effort is necessary to check that all four numbers mentioned above are different."

"When such a number appears as a measure of an angle (expressed as a decimal fraction of seconds) we do not ask ourselves about the probability that a given angle is equal to 13°42′51.7" because we never pose such a question before the measuring. This angle must have some value, and independently of this value we may say after measuring that the a priori probability that this is the value is equal to one divided by ten millions and that this fact is improbable..."

"The question is whether we may say the same if one of the angles in a triangle formed by three stars has a remarkable value, for example, is equal to the angle in an equilateral triangle or ... is equal to half of a right angle ... . In this connection one must say that the tendency to declare an event not specified before the experiment as a remarkable one is very dangerous, because the number of events remarkable from different viewpoints may be very large".

We leave these passages without comment and mention only that it is hard to imagine how the fact that somebody proposed something before the experiment can be taken into consideration in a mathematcial theory.

One more question connected with the application of probability theory is the following. Assume that a statistical hypothesis is chosen. How can we use it? We used to think that the goal of science is to predict something. But probability theory cannot predict anything with certainty; all its predictions have a probabilistic nature. "The vicious circle is apparent ... certainty being impossible, whatever A (the probability axiom) is made to state can only be in terms of 'probability'" (Littlewood [34], pp. 55 – 56]).

We have discussed the difficulties that arise when one tries to apply probability theory to events in the real world. Let us try to point out a way to overcome these difficulties (following [51]).

The application of probability theory has two stages. At the first stage we try to estimate the concordance between statistical hypothesis and experimental results. The rule "the actual occurrence of an event to which a certain statistical hypothesis attributes a small probability is an argument against that hypothesis" ([44], Vol. II, Ch. XIV, part 7, p. 76), it seems, can be made more correct if we are allowed to consider only "simply described" events. It is clear that the event "1000 tails appeared" can be described more simply than the event "a sequence A appeared", where A is a "random" sequence of 1000 heads and tails (these two events have the same probability). This difference may explain why our reactions to these events (we have in mind the hypothesis of a fair coin) are so different. To clarify the notion of a "simply described event" the notion of entropy of the constructive object (introduced by Kolmogorov, see Ch. III) may be useful.

Let us assume that we have already chosen a statistical hypothesis concordant (as we think) with the results of observations. Then we come to the second stage and derive some conclusions from the hypothesis chosen. Here we have to admit that probability theory makes no predictions but can only recommend something: if the probability (computed on the basis of the statistical hypothesis) of an event A is greater than the probability of an event B, then the possibility of the event A must be taken into consideration to a greater extent than the possibility of the event B.

One can conclude that events with very small probabilities may be ignored. In the already cited book [5] Borel writes: "... Fewer than a million people live in Paris . Newspapers daily inform us about the strange events or accidents that happen to some of them. Our life would be impossible if we were afraid of all adventures we read about. So one can say that from a practical viewpoint we can ignore events with probability less than one millionth ... . Often trying to avoid something bad we are confronted with even worse ... . To avoid this we must know well the probabilities of different events" (Russian ed., pp. 159 – 160).

Sometimes the criterion for selection of a statistical hypothesis and the rule for its application are united in the statement "events with small probabilities do not happen". For example, Borel writes "One must not be afraid to use the word "certainty" to designate a probability that is sufficiently close to 1" ([6], Russian ed., p. 7). But we prefer to distinguish between these two stages, because at the first stage the existence of a simple description of an event with small probability is important, and at the second stage it seems unimportant. (We can expect, however, that events interesting to us have simple descriptions because of their interest.)

References

[1] E.A. Asarin, Individual random continuous functions, Proc. 1st World Congress of the Bernoulli Society on mathematical statistics and probability theory, Nauka, Moscow 1986, vol. 1, p. 450.

[2] ———, Some properties of Kolmogorov Δ-random finite sequences, Teor. Veroyatnost. i Primenen. **32** (1987), 556 – 558. MR **89j**:68074.
= Theory Probab. Appl. **32** (1987), 507 – 508.

[3] ———, On some properties of finite objects random in an algorithmic sense, Dokl. Akad. Nauk SSSR **295** (1987), 782 – 785. MR **88m**:68025.
= Soviet Math. Dokl. **36** (1988), 109 – 112.

[4] ——— and A.V. Pokrovskii, Application of Kolmogorov complexity to the analysis of the dynamics of controllable systems, Avtomatik. i Telemekh., **1986**, no. 1, 25 – 33. MR **87e**:93096.
= Automat. Remote Control **47**:1 (1986), 21 – 28.

[5] E. Borel, Le hasard, Alcan, Paris 1914.
Russian translation: *Sluchai*, Gosizdat, Moscow – Petrograd 1923.

[6] ———, Probabilité et certitude, Presses Univ. de France, Paris 1950. MR **12** – 618.
Russian translation: *Veroyatnost' i dostovernost'*, Fizmatgiz, Moscow 1961.

[7] G.J. Chaitin, On the length of programs for computing finite binary sequences, J. Assoc. Comput. Mach. **13** (1966), 657 – 569. MR **35** # 1412.

[8] ———, Incompleteness theorems for random reals, Adv. in Appl. Math. **8** (1987), 119 – 146. MR **88h**:68038.

[9] ———, Algorithmic information theory, Cambridge Univ. Press, Cambridge – New York 1987. MR **89g**:68022.

[10] A. Church, On the concept of a random sequence, Bull. Amer. Math. Soc. **46** (1940), 130 – 135. MR **1** – 149.

[11] A.P. Dawid, Calibration-based empirical probability, Ann. Statist. **13** (1985), 1251 – 1285. MR **87a**:60010.

[12] Yu. L. Ershov, Computable functions of finite types, Algebra i Logika **11** (1972), 367–437. MR **50** # 12688.
= Algebra and Logic **11** (1972), 203–242.

[13] P. Gács, The symmetry of algorithmic information, Dokl. Akad. Nauk SSSR **218** (1974), 1265–1267. MR **53** # 7611.
= Soviet Math. Dokl. **15** (1974), 1477–1480.

[14] ———, On the relation between descriptional complexity and algorithmic complexity, Theoret. Comput. Sci. **22** (1983), 71–93. MR **84h**:60010.

[15] ———, Every sequence is reducible to a random one, Inform. and Control **70** (1986), 186–192. MR **87k**:03043.

[16] K. Jacobs, Turing-Maschinen und zufällige 0-1-Folgen, Selecta Mathematica, Springer, Berlin 1970, vol. II, pp. 141–167. MR **44** # 6413.
Russian translation in: *Mashiny T'yuringa i rekursivnye funktsii*, Mir, Moscow 1972, pp. 183–215.

[17] A.N. Kolmogorov, On tables of random numbers, Sankhyā Ser. A **25** (1963), 369–376.
Russian translation in: *Semiotika i Informatika* No. 18, 1982, 3–13.

[18] ———, Three approaches to the definition of the concept "quantity of information", Problemy Peredachi Informatsii **1**:1 (1965), 3–11. (Reprinted in [22], pp. 213–223.) MR **32** # 2273.
= Problems Inform. Transmission **1**:1 (1965), 1–7.

[19] ———, On the logical foundation of information theory and probability theory, Problemy Peredachi Informatsii **5**:3 (1969), 3–7. (Reprinted in [22], pp. 232–237.)
=Problems Inform. Transmission **5**:3 (1969), 1–4.

[20] ———, Combinatorial foundations of information theory and the calculus of probabilities, Uspekhi Mat. Nauk **38**:4 (1983), 27–36. (Reprinted in [22], pp. 238–250.)
=Russian Math. Surveys **38**:4 (1983), 29–40.

[21] ———, On the logical foundations of probability theory, in: Probability theory and mathematical statistics (Tbilisi 1982), Lecture Notes in Math. **1010**, 1–5.[1] MR **85h**:60007.
Russian translation in: *Teoriya veroyatnostei i matematicheskaya statistika*, Nauka, Moscow 1986, pp. 467–471.

[22] ———, *Teoriya informatisii i teoriya algoritmov* (Information theory and the theory of algorithms), Nauka, Moscow 1987. MR **89a**:01104.

[23] ——— and V.A. Uspenskii, Algorithms and randomness, Teor. Veroyatnost. i Primenen. **32** (1987), 425–455. MR **89d**:68035.
= Theory Probab. Appl. **32** (1987), 389–412.

[24] M. van Lambalgen, Von Mises' definition of random sequences reconsidered, J. Symboic Logic **52** (1987), 725–755. MR **88i**:60005.

[25] ———, Random sequences, Academisch Proefschrift, Amsterdam 1987.

[26] ———, The axiomatization of randomness, Preprint, Univ. of Amsterdam, 1989.

[27] L.A. Levin, The concept of random sequence, Dokl. Akad. Nauk SSSR **212** (1973), 548–550. MR **51** # 2346.
= Soviet Math. Dokl. **14** (1973), 1413–1416.

[1] This text is a reconstruction of A.N. Kolmogorov's speech at the conference made (from a very bad tape recording) by A.K. Zvonkin, A.A. Novikov and A. Shen in the absence of Kolmogorov.

[28] L.A. Levin, Laws on the conservation (zero increase) of information, and questions on the foundations of probability theory, Problemy Peredachi Informatsii **10**:3 (1974), 30–35. MR **57** # 9298.
= Problems Inform. Transmission **10** (1974), 206–210.

[29] ———, Uniform tests for randomness, Dokl. Akad. Nauk SSSR **227** (1976), 33–35. MR **54** # 2325.
= Soviet Math. Dokl. **17** (1976), 337–340.

[30] L.A. Levin, The various measures of the complexity of finite objects (an axiomatic description), Dokl. Akad. Nauk **227** (1976), 804–807. MR **54** # 4185.
= Soviet Math. Dokl. **17** (1976), 522–526.

[31] ———, The principle of conservation of information in intuitionistic mathematics, Dokl. Akad. Nauk SSSR **227** (1976), 1239–1296. MR **58** # 21509.
= Soviet Math. Dokl. **17** (1976), 601–605.

[32] ———, A concrete way of defining measures of complexity, Dokl. Akad. Nauk SSSR **234** (1977), 536–539. MR **56** # 15376.
= Soviet Math. Dokl. **18** (1977), 727–731.

[33] ———, A concept of independence with applications in various fields of mathematics, MIT Computer Science Laboratory Technical Report 235, April 1980.

[34] J. Littlewood, A mathematician's apology, Methuen, London 1953.
Russian translation: *Matematicheskaya smes'*, 4th ed., Nauka, Moscow 1978.

[35] D.W. Loveland, A new interpretation of the von Mises concept of random sequence, Z. Math. Logik Grundlagen Math. **12** (1966), 279–294. MR **34** # 5124.

[36] ———, The Kleene hierarchy classification of recursively random sequences, Trans. Amer. Math. Soc. **125** (1966), 497–510. MR **34** # 7377.

[37] P. Martin-Löf, On the concept of a random sequence, Teor. Veroyatnost. i Primenen. **11** (1966), 198–200.
= Theory Probab. Appl. **11** (1966), 177–179.

[38] ———, The definition of random sequences, Information and Control **9** (1966), 602–619. MR **36** # 6228.

[39] ———, On the notion of randomness, in: Intuitionism and proof theory (Proc. Conf., Buffalo, NY, 1968), North-Holland, Amsterdam 1970, pp. 73–78. MR **43** # 1237.
Russian translation in: *Slozhnost' vychislenii i algoritmov*, Mir, Moscow 1974, pp. 364–369.

[40] ———, Notes on constructive mathematics, Almqvist and Wiksell, Stockholm 1970. MR **58** # 5098.
Russian translation: *Ocherki po konstruktivnoi matematike*, Mir, Moscow 1975.

[41] R. von Mises, Grundlagen der Wahrscheinlichkeitsrechung, Math. Z. **5** (1919), 52–89.

[42] ———, Wahrscheinlichkeit, Statistik und Wahrheit, Springer, Vienna 1928.
Russian translation: *Veroyatnost' i statistika*, Gosizdat, Moscow–Leningrad 1930.

[43] ———, On the foundations of probability and statistics, Ann. Math. Statistics **12** (1941), 191–205. MR **3**–1.

[44] G. Pólya, Mathematics and plausible reasoning. I, Induction and analogy in mathematics, II, Patterns of plausible inference, Princeton Univ. Press, Princeton, NJ, 1954. MR **16**–556.
Russian translation: *Matematika i pravdopodobnye rassuzhdeniya.* I, *Induktsiya i analogiya v matematike*, II, *Skhemy pravdopodobnykh umozaklyuchenii*, Izdat. Inostr. Lit., Moscow 1957.

[45] A. Renyi, Mathematical trilogy [(i) Dialogues on mathematics. (ii) Letters on probability. (iii) Diary—notes of a student in information theory], Russian translation from the Hungarian, Mir, Moscow 1980.
German translation of (i): Birkhäuser, Basel 1967. MR **36** # 7.
English translation of (ii): Wayne State Univ. Press, Detroit, MI, 1972. MR **50** # 5869.

[46] H. Rogers, Theory of recursive functions and effective computability, McGraw-Hill, New York 1967. MR **37** # 61.
Russian translation: *Teoriya rekursivnykh funktsii i effektivnaya vychislimost'*, Mir, Moscow 1972. MR **50** # 4262.

[47] C.-P. Schnorr, Zufalligkeit und Wahrscheinlichkeit. Eine algorithmische Begründung der Wahrscheinlichkeitstheorie, Lecture Notes in Math. **218** (1971). MR **54** # 2328.

[48] ———, Process complexity and effective random tests, J. Comput. System Sci. **7** (1973), 376–378. MR **48** # 3713.

[49] ———, A survey of the theory of random sequences, in: Logic, foundations of mathematics and computability theory, Reidel, Dordrecht 1977, pp. 193–211.

[50] A.Kh. Shen', The frequency approach to the definition of a random sequence, Semiotika i Informatika No. 18, 1982, 14–42.

[51] ———, On the logical basis of applications of probability theory, in: Proc. School-seminar on semiotic aspects of the formalization of intellectual activity (Telavi 1983), VINITI, Moscow 1983, pp. 144–146.

[52] ———, The notion of (α, β)-stochasticity in Kolmogorov's sense and its properties, Dokl. Akad. Nauk SSSR **271 (1983), 1337**–1340. MR **85e**:68034.
= Soviet Math. Dokl. **28** (1983), 295–299.

[53] ———, Algorithmic variants of the notion of entropy, Dokl. Akad. Nauk SSSR **276** (1984), 563–566. MR **86f**:94023.
= Soviet Math. Dokl. **29** (1984), 569–573.

[54] ———, On relations between different algorithmic definitions of randomness, Dokl. Akad. Nauk SSSR **310** (1988), 548–552. MR **90c**:68034.
= Soviet Math. Dokl. **38** (1989), 316–319.

[55] V.A. Uspenskii and A.L. Semenov, *Teoriya algoritmov: osnovnye otkrytiya i prilozheniya* (Theory of algorithms: the main discoveries and applications), Nauka, Moscow 1987. MR **90a**:03057.

[56] Li Ming and P.M.B. Vitanji, Two decades of applied Kolmogorov complexity, Uspekhi Mat. Nauk **43**:6 (1988), 129–166. Russian translation of an article to appear in: Handbook of Theoretical Computer Science (ed. J. van Leeuwen), North-Holland, Amsterdam.

[57] V.G. Vovk, Algorithmic information theory and the problem of prediction, in: *Slozhnostnye problemy matematicheskoi logiki* (Complexity problems of mathematical logic), Kalinin Gos. Univ., Kalinin 1985, pp. 21–24. MR **87a**:68003.

[58] ———, On the concept of the Bernoulli property, Uspekhi Mat. Nauk **41**:1 (1986), 185–186.
= Russian Math. Surveys **41**:1 (1986), 247–248.

[59] ———, On a version of Mises–Church randomness, Proc. 4th All-Union Conf. on application of the methods of mathematical logic, Tallin 1986, pp. 95–97.

[60] ———, On the use of the algorithmic notions of randomness and simplicity in mathematical statistics, Proc. 1st World Congress of the Bernoulli Society on mathematical statistics and probability theory, Nauka, Moscow 1986, vol. 1, p. 456.

[61] ———, The law of the iterated logarithm for Kolmogorov random, or chaotic, sequences, Teor. Veroyatnost. i Primenen. **32** (1987), 456–468. MR **89e**:60072.
= Theory Probab. Appl. **32** (1987), 413–425.

[62] V.G. Vovk, On a randomness criterion, Dokl. Akad. Nauk SSSR **294** (1987), 1298–1302. MR **89a**:60007.
= Soviet Math. Dokl. **35** (1987), 656–660.

[63] ———, On the Kolmogorov–Stout law of the iterated logarithm, Mat. Zametki **44** (1988), 27–37. MR **89k**:60042.
= Math. Notes **44** (1988), 502–507.

[64] V.V. V'yugin, Algorithmic entropy (complexity) of finite objects and its appliation to the definitions of randomness and amount of information, Semiotika i Informatika No. 16, 1981, 14–43.

[65] ———, Some estimates for non-stochastic sequences, Proc. 1st World Congress of the Bernoulli Society on mathematical statistics and probability theory, Nauka, Moscow 1986, vol. 1, p. 455.

[66] ———, On the defect of randomness of a finite object with respect to measures with given complexity bounds, Teor. Veroyatnost. i Primenen. **32** (1987), 558–563. MR **89b**:60014.
= Theory Probab. Appl. **32** (1987), 508–512.

[67] A.K. Zvonkin and L.A. Levin, The complexity of finite objects and the development of the concepts of information and randomness by means of the theory of algorithms, Uspekhi Mat. Nauk **25**:6 (1970), 85–127. MR **46** # 7004.
= Russian Math. Surveys **25**:6 (1970), 83–124.

Translated by A. Shen'

Moscow State University
Scientific Council on the Complex Problem
of Cybernetics, USSR Academy of Sciences
Institute for Problems of Information
Transmission, USSR Academy of Sciences

Received by the Editors 10 July 1989